Universitat Rovira i Virgili

TRIANGLE

LANGUAGE, LITERATURE, COMPUTATION
LENGUAJE, LITERATURA, COMPUTACIÓN
LLENGUATGE, LITERATURA, COMPUTACIÓ

14

[publicacions]
urv

2016 · ISSN 2013-939X · DOI 10.17345/triangle
https://revistes.urv.cat/index.php/triangle

REVISTA TRIANGLE

Directores · Editors-in-Chief
Dra. M. Dolores Jiménez López, Universitat Rovira i Virgili, Tarragona (Spain)
Dr. Antonio M. García Español, Universitat Rovira i Virgili, Tarragona (Spain)

Consejo editorial · Editorial Board
Leonor Becerra Bonache, Université Jean Monnet, Saint Etienne (France)
Natalia Català Torres, Universitat Rovira i Virgili, Tarragona (Spain)
Erzsébet Csuhaj-Varjú, Eötvös Loránd University, Budapest (Hungary)
Elena de la Cruz Vergari, Universitat Rovira i Virgili, Tarragona (Spain)
Carlos Martín Vide, Universitat Rovira i Virgili, Tarragona (Spain)
Alfonso Ortega de la Puente, Universidad Autónoma de Madrid, Madrid (Spain)
M. José Rodríguez Campillo, Universitat Rovira i Virgili, Tarragona (Spain)
György Vaszil, University of Debrecen, Debrecen (Hungary)

Comité científico · Scientific Committee
Dulce María Bautista, Universidad Central, Bogotá (Colombia)
Antoni Brosa Rodríguez, Uniwersytet Łódzki, Lodz (Poland)
M. Isabel Calle Romero, Universitat Rovira i Virgili, Tarragona (Spain)
Emmanuele Chersoni, Hong Kong Polytechnic University, Hong Kong (China)
Henning Christiansen, Roskilde University, Roskilde (Denmark)
Dolores Corbella Díaz, Universidad de La Laguna, Tenerife (Spain)
Anca Crivat, University of Bucharest, Bucharest (Romania)
Veronica Dahl, Simon Fraser University, Burnaby (Canada)
Adrian-Horia Dediu, University Politehnica of Bucharest, Bucharest (Romania)
Marina de la Cruz Echeandía, Universidad Autónoma de Madrid, Madrid (Spain)
Esther Forgas Berdet, Universitat Rovira i Virgili, Tarragona (Spain)
José Eugenio García Albea, Universitat Rovira i Virgili, Tarragona (Spain)
Adela Grando, San Diego Super Computer Center, San Diego (USA)
M. del Carmen Horno Chéliz, Universidad de Zaragoza, Zaragoza (Spain)
Iraide Ibarretxe Antuñano, Universidad de Zaragoza, Zaragoza (Spain)
Vincenzo Manca, Università di Verona, Verona (Italy)
Antonio Moreno Ribas, Universitat Rovira i Virgili, Tarragona (Spain)
Benedek Nagy, University of Debrecen, Debrecen (Hungary)
Isabel Oltra Massuet, Universitat Rovira i Virgili, Tarragona (Spain)
Diana Pérez Marín, Universidad Rey Juan Carlos, Madrid (Spain)
Łukasz Szkopiński, Uniwersytet Łódzki, Lodz (Poland)
Cristina Tîrnăucă, Universidad de Cantabria, Santander (Spain)
Adrià Torrens Urrutia, Universitat Rovira i Virgili, Tarragona (Spain)
Florentina Lilica Voicu, Universitat Rovira i Virgili, Tarragona (Spain)
Michael Zock, LIF-CNRS, Marseille (France)

ISSN: 2013-939X · DL: T-1492-2010 · DOI 10.17345/triangle
http://revistes.urv.cat/index.php/triangle

Edita · Publishing
Publicacions de la Universitat Rovira i Virgili:
Av. Catalunya, 35 - 43002 Tarragona
Tel. 977 558 474 · publicacions@urv.cat

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-nd/4.0/

TRIANGLE. *Language, Literature and Computation*
Volume 14, 2016
DOI: 10.17345/triangle14
ISSN: 2013-939X
http://revistes.urv.cat/index.php/triangle

Volume Contributors

Ife Adebara
Department of Computing Science
Simon Fraser University
iadebara@sfu.ca

Susana Campillo Muñoz
Departament de Filologies Romàniques
Universitat Rovira i Virgili
susana.campillo@urv.cat

Contents

Using Womb Grammars for Inducing the Grammar of a Subset of Yorùbá Noun Phrases
IFE ADEBARA 1

Violencia Lingüística en las Redes Sociales
SUSANA CAMPILLO MUÑOZ 37

TRIANGLE. *Language, Literature and Computation*
Volume 14, 2016, pp. 1–36
DOI: 10.17345/triangle14.1-36
ISSN: 2013-939X
http://revistes.urv.cat/index.php/triangle

Using Womb Grammars for Inducing the Grammar of a Subset of Yorùbá Noun Phrases

IFE ADEBARA*

Simon Fraser University, Burnaby, Canada

Abstract. We address the problem of inducing the grammar of an under-resourced language, Yorùbá, from the grammar of English using an efficient and, linguistically savvy, constraint solving model of grammar induction –Womb Grammars (WG). Our proposed methodology adapts WG for parsing a subset of noun phrases of the target language Yorùbá, from the grammar of the source language English, which is described as properties between pairs of constituents. Our model is implemented in CHRG (Constraint Handling Rule Grammar) and, it has been used for inducing the grammar of a useful subset of Yorùbá Noun Phrases. Interesting extensions to the original Womb Grammar model are presented, motivated by the specific needs of Yorùbá and, similar tone languages.

Keywords: property grammars, constraint handling rules grammar, constraint handling rules, womb grammar, computational grammar, syntax, grammar induction, Yorùbá.

1 Introduction

Grammar induction is a data driven approach to learning grammars, and, analyzing languages.

* **Author's address**:
Department of Computing Science
Simon Fraser University
8888 University Drive, Burnaby, Canada
E-mail iadebara@sfu.ca

We adopt a novel grammar induction technique, called, Womb Grammars (WG), developed by Dahl and Miralles (2012a). WG employ a property based methodology, which, relies on the grammar of a source language to induce the grammar of a target language. WG assume that the grammar of the source language is correct and, that the lexicon and, input phrases of the target language are correct and, representative of a fragment of noun phrases in the target language. It adopts a constraint-satisfaction approach whereby input phrases of the target language are tested for satisfaction and, unsatisfied constraints provide a lead for the reconstruction of the target grammar using the source grammar. We use a context free grammar of the target language to evaluate the correctness of our parser.

In this paper, WG are used to induce a/the grammar of a subset of noun phrases in Yorùbá, from that of an English grammar with same coverage (Dahl & Miralles 2012b). We use English as our source language and, Yorùbá as the target language. English was chosen as the source language for various reasons, especially because, it is a language the author can speak and, write fluently. English also has a wide audience, and, we believe that it will be easier for a wider audience to understand the WG model using a language: English, whose structure is well known. Using a well known language eliminates the need for the reader to in addition to understanding the WG model, learn and, understand the structure of two less known languages (Source and target languages). Also, because English is a language which has been well studied by linguists, many linguistic resources are available and, easily accessible, and, this made creating a correct property grammar (details in section 4.1) easier than using any language which has less resources easily accessible. The author's formal training in linguistics and, native speaker competence in Yorùbá also made Yorùbá a desirable language as the target. Since we need a correct grammar of both the source and, target language, English and, Yorùbá languages were thus, the most accessible languages for this task. In fact, the success achieved from using English to induce the grammar if Yorùbá, reveals that, it is possible to use WG for any pair of languages, including unrelated languages, although with some modifications.

The novel approach of WG, needs neither a pre-specified model family, nor, parallel corpora, nor any of the typical models of machine learning, and, works for more than just specific tasks such as disambiguation (Dahl & Miralles 2012a). It automatically transforms a given (task-independent) grammar description of a source language, from just the lexicon of the target language, and, a representative input set of correct phrases in the target language. The syntactic descriptions of the source and, the target language subset addressed are stated in terms of linguistic constraints (also called "properties" in linguistic literature) between pairs of constituents, although for the

target language constraints we depart from the classic formalism (Blache 2004) in view of Yorùbá motivated extensions. This allows us a great degree of modularity, in which constraints can be checked efficiently through constraint solving.

Using linguistic information from one language to describe another language has yielded good results, however, it was used for tasks like disambiguating the other language (Burkett & Klein 2008), fixing morphological or syntactic differences by modifying tree-based rules (Nicolas *et al.* 2009), and, not for syntax induction which is our focus. The approach adopted usually requires parallel corpora, although, there exists an exception (Cohen & Smith 2010) where a bilingual dataset is used to train parsers for two languages simultaneously. This is accomplished by tying grammar weights in the two hidden grammars, and, is useful for learning dependency structure in an unsupervised empirical Bayesian framework.

Our results in applying and, adapting the WG framework for inducing Yorùbá noun phrases show that this model compares favourably with others in solving the grammar induction problem: it combines linguistic formality with efficient problem solving, and, can transfer into other languages, including languages in which tones have a grammatical and, or, semantic function.

2 Motivation

Language endangerment and, death has been of serious concern in linguistics and, language policy making. Close to seven thousand languages are currently spoken in the world, the majority of which are understudied and, endangered. It has been said that an alarming 50 to 90 percent of languages will be extinct by the end of the century (Romaine 2002).

For various reasons, some speakers of many minor, less studied languages may learn to use a different language from their mother tongue and, may even stop using their native languages. Parents may begin to use only that second language with their children and, gradually the transmission of the native language to the next generation is reduced and, may even cease. As a result, only the elderly in such communities may use the native language, after a while, there may be no speakers who use the language as their first or, primary language and, eventually the language may no longer be used at all. Thus, a language may become extinct, existing perhaps only in voice recordings, written records and, transcription and, languages which have not been adequately documented completely disappear.

Linguists cannot keep up with the study of the endangered languages even for educational purposes, and, there is a growing need for their automatic processing as well, since the amount of text sources grows much faster than humans can process them. To make matters worse, most linguistic resources support English and, a handful of other "first world" languages, leaving the vast majority of languages and, dialects under-explored. Clearly, automating the discovery of an arbitrary language's grammar model would render phenomenal service to the study and, preservation of linguistic diversity.

Scientifically, we explore to what extent the parsing-as-constraint-solving paradigm of natural language processing problem solving could achieve a great degree of linguistic, descriptive formality without sacrificing efficiency, in the realm of grammar induction, and, in particular for inducing Yorùbá, which is severely under-resourced and, endangered.

3 The Yorùbá Language

Yorùbá belongs to the Yoruboid group of the Kwa branch of the Niger-Congo language family, which cuts across most of sub-Saharan Africa. It is a tonal dialect-continnum comprising about 20 distinctive dialects and, spoken by over 30 million people in the western part of Nigeria (Fagborun 1994). Niger-Congo is the largest of the five main language families of Africa. The others being Nilo-Saharan, Afro-Asiatic, Khoisan and, Austronesian (mainly found in the nation of Madagascar).

Yorùbá is one of the three regional (national language contained in the constitution) languages in Nigeria and, is said to be the most studied African language. Yorùbá is spoken by more than 20 percent of 170 million people who make up the population of Nigeria (the largest single black nation on earth). The two other national languages are Hausa and, Igbo, both of which are also regional languages in the north and, southeastern parts of the country respectively. The Yorùbá language is a koine (Fagborun 1994), a process involving dialect mixing, levelling, and, simplification (Trudgill 1986). Standard Yorùbá which can also be referred to as the standard koine is a compendium of several dialects, mainly dialects of Ọ̀yọ́, Ìbàdàn, Abẹ́òkúta, and, Lagos, which were major trade centers of early Church Mission Society missionary activities.

Apart from the standardized koine, there are twenty other dialects of Yorùbá: Ọ̀yọ́, Ìjẹ̀ṣà, Ìlá, Ìjẹ̀bú, Oǹdó, Wo, Owe, Jumu, Iworro, Igbonna, Yagba, Gbedde, Ẹ̀gbá, Akono, Aworo, Bunu (Bini), Èkìtì, Ìlàjẹ, Ìkálẹ̀, Awori. These dialects are largely

mutually intelligible, albeit with some variations in vocabulary and, phonology and, were largely spoken by different groups of people who, though tracing their descent to their common progenitor (Oduduwa), do not consider themselves as one people. Although there are several dialects of Yorùbá, it is important to mention that the present research is based on the 'standard' Yorùbá. This is because, 'standard' Yorùbá is the only variety referred to as Yorùbá and, the only variety taught in schools in Yorùbá land and, diaspora. It is the language of the media and, of official government business. It is described as a resource scarce language (Adeyanju *et al.* 2015) due to the limited number of texts available in Yorùbá.

3.1 The Sociolinguistic Situation of Yorùbá in Nigeria

Despite the seemingly large population of Yorùbá speakers, according to Bámgbósé (1993) it is a deprived language. Yorùbá has also been classified as a seriously endangered language (Fabunmi & Salawu 2005). This is according to Wurm (1998) who categorises language status into:

(i) potentially endangered - there is a decline in the use of the language by children
(ii) endangered - youths have experienced a decline in the use of the language and children experience language loss
(iii) seriously endangered - competent speakers are about 50 years old and, above
(iv) moribund - competent speakers are very old, and, many are dead
(v) extinct - when the last speaker dies;

The status of Yorùbá is influenced by the language policy of Nigeria, which favours the use of English above all indigenous languages. It also pays lip service, for instance, to the national language policy of education, which states that the mother tongue or, the language of the immediate community must be adopted as the language of education in primary schools, and, English should only be introduced at a later stage. Other language policies in other domains that encourage the use of mother tongues are also not adopted. So that the regional or, national status of Yorùbá and, other regional languages is theoretically, but not fully implemented in practice (Fabunmi & Salawu 2005).

Colonialism imported the English language to Nigeria and, English has since been adopted as the official language in government, education, and, all official businesses. English is the language of the elite, and, fluency in English is synonymous with a good education. As a result, many parents, even those who are barely educated or, not educated at all, ensure that their children are taught in English right from the

elementary classes. In most schools, indigenous languages are referred to as vernacular and, are prohibited. Violation usually attract fines and, many times corporal punishment.

Bilingualism is also believed to affect children's ability to attain competence in English, and, thus, parents avoid speaking mother tongues at home for fear of raising children with poor communication in English. Many children therefore can neither speak, read nor write in Yorùbá, and, many do not even understand the language at all.

In addition, the great linguistic diversity in Nigeria discourages people from speaking Yorùbá as often as they may have. Well over 450 languages are said to be spoken in Nigeria, therefore necessitating the use of English as the lingua franca (Less educated people use the Nigerian Pidgin (an English based creole of some sort) as a medium of communication, and, some educated people use the Nigerian pidgin in less formal environments) (Adegbija 2004).

English is also the language spoken in offices, corporate organizations, and, the use of Yorùbá has greatly declined, resulting in lower competency among many users, code-mixing, code switching, and, several other language loss trends. In addition, although Yorùbá is still widely used in media and, there exist radio and, television programs in Yorùbá, as well as newspapers, they receive their patronage from the older generation since there always exist the English versions, which receive a greater patronage.

All the afore mentioned language situation have influenced the lack of adequate language development, and, consequently resulted in, resource scarcity of Yorùbá.

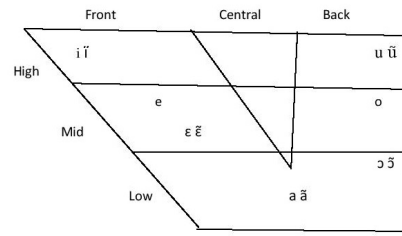
3.2 Yorùbá Phonology

The phonology of Yorùbá, which is how phonemes (Unit of distinctive sound that account for meaning distinction) are strung together to form words, comprises of seven oral vowels, five nasal vowels, three syllabic nasals and, seventeen consonants. Oral vowels are produced from the oral orifice, while nasal vowels are produced through the nasal orifice. Nasalized vowels, those vowels produced with the nasal and, oral orificies, are known to occur through an assimilation process when vowels preceed nasal consonants. All the seven oral vowels can be nasalized.

Pitch, the perceptual correlate of vocal cords' frequency, is distinctive in Yorùbá. This distinctive/phonemic pitch is called tone. For example, the word *rá*(H) (disappear), *rà*(L) (buy) and, *ra*(M) (rub) are distinguished by the three tones in Yoruba:

High-tone, Mid-tone and Low-tone. High-tone is orthographically represented with acute accent, the Low-tone with grave accent and, the mid is unmarked. These tones can interact to form rising and, falling tones for instance *yìí* (this) and, *náà* (the). Vowels and, syllabic nasals are the tone bearing units in Yorùbá.

Consonant clusters are not permitted in Yorùbá but long vowels are possible.



		Labial	Alveolar	Palatal	Velar	Labiovelar	Glottal
stops	Voiceless		t		k	kp	
	Voiced	b	d		g	gb	
Affricates	Voiced			dʒ			
Fricatives	Voiceless	f	s	ʃ			
Nasals		m	n	ŋ			
Approximants	lateral		l				
	Central		r	j		w	h

Fig. 1. Yorùbá Vowel and Consonant Chart.

3.3 Yorùbá Morphology

Typologically, Yorùbá is an analytic language, that is, grammatical relations are expressed with little or, no use of inflection. Tense is not marked morphologically, and, agreement relation is limited. It is a combination of isolating, that is, it has a low morpheme per word ratio, and, agglutinating through the use of prefixes, interfixes, reduplication (full and, partial reduplications) and, compounding. Prefixation and, reduplication are the most common word formation processes in Yorùbá. Word formation processes in Yorùbá are also derivational and, words changed through inflections are usually through tonal variations (Awoyale 1988). Due to the syllable structure of Yorùbá, compounding usually involves vowel ellision, coalescense, epenthesis, and, compensatory lengthening to resolve vowel hiatus and, consonant clusters.

3.4 Yorùbá Orthography

The Yorùbá language has been written since the early 19th century. However, there have been several changes in the orthography since then (Fagborun 1989; Olumuyiwa 2013).

Orthographically, the Yorùbá alphabet consists of 25 letters and, uses the familiar Latin characters. Nasalized vowels are represented with a 'n' after the vowel. Syllabic nasals are written as 'n', except when they occur before 'b', where they are represented as 'm'.

Although tones occur on syllables, they are orthographically marked on vowels and, syllabic nasals. Yorùbá syllables are open; this means they end in vowels. The syllables are formed by a single vowel, consonant plus vowel, or, a syllabic nasal.

The present standards were established in 1974, however, there remains a great deal of contention over writing conventions, spelling, grammar, the use of tone marks and, some other linguistic formalities. For the purpose of this paper, we have adopted the generally accepted formalities of the standard language according to grammars by Bámgbósé (1966) and Awobuluyi (1978). These include full tone marking including tone marks on syllabic nasals.

Aa	**Bb**	**Dd**	**Ee**	**Ẹẹ**	**Ff**	**Gg**	**Gbgb**
[a]	[b]	[d]	[e]	[ɛ]	[f]	[g]	[gb]
Hh	**Ii**	**Jj**	**Kk**	**Ll**	**Mm**	**Nn**	**Oo**
[h]	[i]	[ʤ]	[k]	[l]	[m]	[n] [ŋ]	[o]
Ọọ	**Pp**	**Rr**	**Ss**	**Ṣṣ**	**Tt**	**Uu**	**Ww**
[ɔ]	[kp]	[r]	[s]	[ʃ]	[t]	[u]	[w]
Yy							
[j]							

Fig. 2. The Yorùbá Alphabet and the Sounds they Represent.

3.5 Constituent Structure of Yorùbá Noun Phrases

As in English, the canonical word order is SVO, e.g, as in ọmọ ra oúnjẹ (child buy food). However, there are several interesting differences between the constituent structure of Yorùbá and, English. Five of these are within the scope of noun phrases

and, have been handled in this work. Details of how we address these differences are in section 7.

(1) In the arrangement of constituents, Yorùbá has a determiner final position in structure and, determiners are not obligatory in noun phrases (Ajiboye 2006). So that *ajá kékeré* (dog small) and, *ajá kékeré kan* (dog small a) are felicitous in Yorùbá.

(2) Yorùbá has a superstructure, from which 'adjectives' are derived (Awobuluyi 1978; Bámgbósé 1966; Bowen 1858; Awoyale 2008; Ogunbowale 1970). For instance the word *pupa* (red) has two entries in the dictionary (Awoyale 2008). The first entry is as a noun in *Pupa dára ju dúdú lọ* (red good more black than) which means red is better than black, and, as a verb in *ọmọ́ pupa* (child red) which means, child is red. It is important to understand that in the example given for the first entry, *pupa* and, *dúdú* are nouns because both words can be replaced with the pronoun *'ó* 'he/she/it', so that we can say *ó dára ju ú lọ*. The *ó* is derived as *ú* after *ju* through an assimilation process because Yorùbá abhors vowel hiatus. In cases of plural and, honorific nouns (just like *tu* and, *vous* in French), we will have *Pupa dára ju àwọn dúdú lọ* (red, good more plural, black than), *Pupa dára ju wọ́n lọ* (red good more them (or singular honorific) than)'. In the second example however, *ọmọ́ pupa* has also gone through a vowel deletion and, elongation through an assimilation process, which turned *ọmọ ó pupa* to *ọmọ́ pupa*. The mid tone *o* was deleted and, the high tone *o* was elongated. Although the *ó* (which is a pronoun but functions as a determiner) in that sentense is implied but will not be articulated by native speakers of Yorùbá. *ọmọ pupa* 'red child' however is an example where *pupa* performs a modifier or, adjective function to the noun *ọmọ*. Several other words like "kékeré" 'small', *ńlá* 'big', *púpọ̀* 'plenty' are examples of some nouns or, verbs that can function as adjectives. For the purpose of this paper, nouns or, verbs that modify other nouns are referred to as adjectives.

(3) In Yorùbá, agreement is morphologically absent and, as a result, plurality is shown syntactically through the occurrence of a pronoun *àwọn* 'they' before a noun as in *àwọn ọmọ* (they child) which means 'children'; addition of cardinals or, adjectives after a noun as in *ọmọ méjì* (child two) implies 'two children', addition of an elided form of the pronoun *àwọn*, before a demonstrative as in *ọmọ wọ̀n yẹn* (child they), that means 'those children' or, reduplication of certain 'adjectives' after the noun as in *ajá dúdú dúdú* (dog black black) (Ajiboye 2006). It is important to note that the words used for plurality in Yorùbá have their own distinct meanings and, are used in certain contexts to depict plurality.

For instance *Ilé ńlá ńlá* (house big big) will denote several houses but *esè kòǹgbà kòǹgbà* (leg big big) will not be a plural expression despite both phrases having identical syntactic categories (Ogunbowale 1970).

(4) Yorùbá doesn't have a formal feature for indicating gender and, when gender distinction is mentioned in Yorùbá grammar, it is a translation equivalence, or, it is with reference to the structure of English and, not Yorùbá (Bámgbósé 1966). The use of words like *akọ* (male), *abo* (female), *obìnrin* (female human being), *ọkùnrin* (male human being) can be used to express gender in words like: *Abo kìnìhún* (lioness), *ọmọ obìnri* (child female) or, 'daughter', *akọ ẹlẹ́dẹ̀* (pig) while the pronoun "o" is used to refer to 'he/she/it' (Bámgbósé 1966). We explain how we use gender features in section 8.7.

(5) Another interesting distinction between Yorùbá and, English structure is the position of adjectives in phrases. Many adjectives occur after nouns, but it is also common to have adjectives before the noun. For instance, *òkú* (dead) always occurs before a noun as in *òkú ẹran* (dead meat). Adjectives that describe a part of something always occur before a noun as in *ìdajì oúnjẹ* 'half food'. Adjectives that describe the attribute of a person or, thing usually occurs before a noun as in *kékeré ẹkùn* (small leopard), but there are some exceptions where having the adjective after the noun are accepted as correct (Ogunbowale 1970). We describe our results, which are consistent with this linguistic claim in section 9.

Although, we cannot discount language universalities, these exceptions which can cause difficulty in simple comparisons have been managed in very interesting ways.

4 Linguistic Background

4.1 Data Collection

Data collection for this research has been a combination of introspective and, empirical collection methods. The collection process commenced with the development of a context free grammar of noun phrases in Yorùbá. This grammar was developed by linguists who also have native competence in Yorùbá. The first context free grammar was developed by the author who is also a linguist and, the grammar was extended by the other linguists Dr. Tunde Adegbola (African Languages Technology Initiative), Dr. Demola Lewis (University of Ibadan), Samuel Akinbo (University of British Columbia). The grammar was then used to generate phrases which were used as input phrases in our Womb Grammars model. In all, fourty five phrases were generated. Native speakers of Yorùbá who do not have formal linguistic training were

invited to check the phrases generated by the context free grammar in order to ascertain that those phrases are intuitively correct to the regular user of the language.

We developed a context free grammar because there is no known treebank of Yorùbá and, Penn Treebank only has a dictionary of words.

This approach of data collection was employed in order to ensure that our model is realistic, correct, and, robust. Introspection has proven to be the most reliable process of data collection and, also very useful for building models which require high level linguistic competence such as this WG model (Chomsky 1957). Introspection as previously mentioned, has been contributed by the author who is a native speaker of Yorùbá and, also has formal training in linguistics. We have double-checked our introspective conclusions by consulting seven other native speakers of Yorùbá, three of which also have formal graduate level training in linguistics.

Data collected have also been compared with two existing grammars of Yorùbá. The first by Ayò Bámgbósé (1966) and, the other one by Awóbùlúyí (1978). It was important to observe these existing grammatical descriptions of Yorùbá, considering that they are among the earliest contributions of native speakers who have formal linguistic training to the description of the Yorùbá grammar (Chelliah & de Reuse 2010). Ajiboye's (2006) description of Yorùbá noun phrases has also been very useful for verifying the grammar induced by our model, especially in relation to plural marking.

4.2 Strategies for Part of Speech Parsing

Accuracy of part-of-speech tagging is a critical and, fundamental building block for many computational linguistic tasks including grammar induction. Assigning correct part-of-speech tags to each input word explicitly indicates some inherent grammatical structure of any language and, a wrong part-of-speech tag will distort the grammatical structure of a language. Rule-based, data driven and, hybrid methods for part-of-speech tagging have been extensively described in literature (Wang & Li 2011). Rule-based methods are those derived from linguistic rules (Brill *et al.* 1990), data-driven methods are derived from statistical analysis of language data (Merialdo 1994), and, hybrid methods are a combination of rule driven and, data driven approaches (Sun & Bellegarda 2011).

Data-driven approaches have yielded very good results, albeit for Indo-European languages and, other data rich languages like English, Dutch, German, etc. However, data driven approaches have error rates which are usually reducible by only a few

percentage, and, also work poorly with languages having a less fixed word order (Voutilainen 2012).

Rule-based part-of-speech tagging on the other hand is the oldest approach that uses hand-written rules for tagging. Rule based taggers depend on dictionaries or, lexicons, that contain word forms together with their associated part-of-speech labels, as well as context-sensitive rules to choose the appropriate tags during application. The dictionary or, lexicon is consulted to get possible tags for each word. Hand-written context-sensitive rules are used to identify the correct tag if a word has more than one possible tag. The linguistic features of the word and, other words surrounding it are analyzed for disambiguation purposes. For example, a tagger for English will have a rule such as: if the preceding word in a phrase is a determiner then the last word in the phrase must be a noun.

We adopt a rule-based approach. Rule-based approach, though rigorous and, requiring a great amount of high level linguistic skills, yield good results for any language, including those like Yorùbá, which we have identified as resource scarce in section 3 and, 3.1, as well as having a less fixed word order structure in section 3.4.

The tagsets were developed, with the use of Awoyale (2008), of Yorùbá as well as native speakers of Yorùbá who have formal linguistic training. The tagsets were compared to the grammars of Ayò Bámgbósé (1966) and, Awóbùlúyí (1978), one of the earliest grammar descriptions of Yorùbá by native speakers who have formal linguistic training.

We use the following tagsets:

i noun - ajá (dog), ẹran (goat), etc
ii pronoun - àwọn (they), ìwọ (you), etc
iii proper-noun - Ayò, Bámgbósé, etc
iv determiner - kan (a), náà (the), etc
v quantifier - gbogbo (every), ìdajì (half), etc
vi adjective - dúdú (black), pupa (red), etc

We further define features for each word in order to provide a fine-grained definition to each word tag. We use the following features:

i Number
ii Gender
iii Tone

iv Person
v Definitiveness
vi Case

These features have been carefully chosen to ensure that our model accounts for the unique traits of Yorùbá.

5 Property Grammars

The idea of constraint is present in modern linguistic theories such as Lexical Functional grammars (LFG) and, Head-driven Phrase Structure grammars (HPSG) (Pollard & Sag 1994). However, constraint-satisfaction, a way of implementing constraints, is not really incorporated in the implementation of these theories. We use a formalism called Property Grammar (PG) (Blache 2004), which is based completely on constraints: all linguistic information is represented as properties of pairs of constituents, which allows parsing to be implemented as a constraint-satisfaction problem. The set of properties forms a system of constraints expressed over categories. They represent different kinds of information (e.g. agreement, morphology or, semantics) that can be used typically for contextual restrictions. PG differs from HPSG and, LFG in that it does not belong to the generative syntax family, but to the model-theoretic syntax one (Blache 2000).

Constraints represent information and, the first benefit of representing a problem using constraints is that of partially solving it, so this allows for partial solutions to be found even when a full solution is not available. These constraint-based or, property-based theories, such as Property Grammars (PG) (Blache 2004) evolved from Immediate Dominance/Linear Precedence (IDLP), which unfolds a rewrite rule into the two constraints: (1) of immediate dominance (expressing which categories are allowable daughters of a phrasal category), and, (2) linear precedence (expressing which of the daughters must precede which others).

For example in the PG framework, English noun phrases can be described using constraints such as: precedence (a determiner must precede a noun, an adjective must precede a noun), requirement (a singular non-generic noun must be used with a determiner), obligation (a noun phrase must contain the head noun), and, so on. The linguistic structure (e.g phrase, sentense, clause etc) is characterized by a finite list of the constraints it satisfies and, a list of constraints it violates, so that even incorrect or, incomplete phrases will be parsed. For instance, the phrase the professor

emeritus, which be characterized as satifying obligation, requirement, determiner precedence but will not satify the adjective precedence. This differs from traditional parsers which characterize a linguistic structure with either a parse tree or, a failure. Using PGs also enables us relax certain constraints by declaring conditions under which those constraints should be relaxed. For instance, the obligation rule of English can be relaxed to allow pronouns and, proper nouns instead of only a noun because pronouns and, proper nouns can also function as the head of a phrase. Dahl and Blache (2004) encode the input PG into a set of CHRG rules that directly interpret the grammar in terms of satisfied or, relaxed constraints and, a syntactic tree is the implicit result. Womb Grammars which is an adaptation of this framework into grammar transformation (Dahl & Miralles 2012b; Christiansen 2005) induces a language's syntactic structure from that of another.

Presently, the PG formalism comprises the following seven categories (we adopt the handy notation of Duchier *et al.* (2013), and, the same example here):

Constituency $A : S$, children must have their categories in the set S
Obligation $A : \triangle B$, at least one B child
Uniqueness $A : B!$, at most one B child
Precedence $A : B \prec C$, B children precede C children
Requirement $A : B \Rightarrow C$, if B is a child, then also C is a child
Exclusion $A : B \not\Leftrightarrow C$, B and, C children are mutually exclusive
Dependency $A : B \sim C$, the features of B and, C are the same

Example 1. For example, if we denote determiners by D, nouns by N, proper nouns by PN, verbs by V, noun phrases by NP, and, verb phrases by VP, the context free rules $NP \rightarrow D\ N$ and, $NP \rightarrow N$, which determine what a noun phrase is, can be translated into the following equivalent constraints: $NP : \{D, N\}$, $NP : D!$, $NP : \triangle N$, $NP : N!$, $NP : D \prec N$, $D : \{\}$, $N : \{\}$.

In the PG formalism, A refers to the phrase or, sentence being tested. In our case, A refers to Noun phrases, B and, C are categories such as nouns, adjectives, etc. These properties contain the basic syntactic information. Other properties can be added if necessary, in particular for integrating knowledge coming from other linguistic domains or, for particular phenomena, like long distance dependencies.

The use of a dependency property makes it possible to express a dependency grammar using the formalism of property grammars. This is possible because, just like dependency grammars, dependency property is concerned with syntactico-semantic

relations. For instance np(dependence(adjective,noun)) is concerned with using features (described in section 4.2) to determine the gender and, number relationship between adjectives and, nouns, and, other dependency information can be modelled in similar fashion. The features are introduced for each word in the lexicon using the CHRG symbol word/3 (details in section 8.2). However, all the subsets of categories involved in the description of a given input can be characterized by the constraint system (i.e. the grammar). The constituency property contains dependency information that tell us each category can be associated to another.

Thus, there is a general notion of characterization formed by the set of properties that succeed and, those that fail which together characterize an input. The characteristics of an input corresponds to the result derived from the constraint system at the end of the parse. The result of a parsed property is either satisfied or, unsatisfied. The interest of such a conception is that every property in the grammar states some constraint on well formedness. It can be the case that all constraints are satisfied, but this is not an imperative condition. All kinds of input can consequently receive a characterization.

6 Background on Womb Grammars

6.1 Womb Grammars

Womb Grammar Parsing was motivated by the need to aid the world's linguist in describing the syntax of the many languages that are not being studied for lack of adequate resources. It induces the grammar of a target language from the grammar of a source language. The WG paradigm describes a language's phrases in terms of constraints over properties of pairs of direct daughters called properties. WG extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing.

WG was presented in two versions: (1) *Hybrid Womb Grammars*, in which the source language is an existing language for which the syntax is known, and, (2) *Universal Womb Grammars*, in which the source syntax is a hypothetical universal grammar of the authors' own devise, which contains all possible properties of pairs of constituents. Womb Parsing has the novel approach of addressing a problem which more usually is solved as a machine learning problem through constraint solving.

The grammar of our source language, English is described as properties of constituents as described by Blache (2004). We use the hybrid version of Womb Grammars, to which we feed the following English grammar:

Constituency *NP* : *determiner, noun, adjective, pronoun, proper-noun, quantifier*;
Obligation *NP* : $\triangle$*noun*; $pronoun; proper - nouns$
Precedence *NP* : *determiner* $\prec$ *noun*; *NP* : *determiner* $\prec$ *adjective*; *NP* : *pronoun* $\prec$ *noun*; *NP* : *pronoun* $\prec$ *adjective*; *NP* : *adjective* $\prec$ *noun*; *NP* : *quantifier* $\prec$ *determiner*; *NP* : *quantifier* $\prec$ *pronoun*; *NP* : *quantifier* $\prec$ *adjective*; *NP* : *quantifier* $\prec$ *noun*
Requirement *NP* : *noun* $\Rightarrow$ *determiner*
Dependency *NP* : *quantifier* $\sim$ *noun*; *NP* : *adjective* $\sim$ *noun*; *NP* : *determiner* $\sim$ *noun*

The nouns used in the noun phrases are proper nouns, common nouns, abstract nouns, pronouns etc and the features such as number, gender, etc are encoded in the features of the word/3 (more details in section 4.2 and 8.2). WG extends the parsing capabilities implicit in these properties into a model of grammatical induction, in addition to parsing. The first implementation of WG (Dahl & Miralles 2012b) allowed greater efficiency by adopting a failure-driven approach where only failed constraints where analyzed, rather than explicitly calculating all successful and, unsuccessful constraints. However in adapting the model into Yorùbá, we have found it more efficient to explicitly calculate satisfied constraints as well. This is partly because in order to arrive at as nuanced a description as needed for Yorùbá, we had to extend the Property Grammar model to accommodate what we call **conditional constraints**: those constraints which have failed in some phrases and, succeeded in others, but also have a unique pattern responsible for this behaviour (more on this later).

The general WG model can be described as follows:

Let L^T be the target language and, its lexicon (L^T_{lex}). Let E^T be a given expression (a meaningful and, grammatical expression) of a target language L^T. Likewise, let L^S be the source language. Its syntactic component will be denoted L^S_{syntax}. Let Seq^S be a sequence of words of L^S, which is a glossing correspondence i.e a literal transliteration (does not need to be meaningful and, or, grammatical) of E^T and, E^S be an expression of the language L^S which is a meaningful translation of E^T. The translation expression E^S in L^S of E^T in L^T should be meaningful and grammatical. If we can get hold of a sufficiently representative set of phrases, E^T, in L^T that are known to be correct (a set in which our desired subset of the target language will be represented), we can feed these to a hybrid parser consisting of L^S_{syntax} E^T, and, L^T_{lex}. This will result in some of the sentences being marked as incorrect by the parser. An analysis of the constraints these "incorrect" sentences violate can subsequently reveal how to transform L^S_{syntax} so that the parser accepts as correct the

sentences in the corpus of L^T—i.e., how to transform L^S_{syntax} into L^T_{syntax} by modifying the constraints that were violated into constraints that accept the input. Seq^S and E^S are thus the gloss and translation expression of E^T, incoporated to enable intelligibility for non-speakers of the target language. Figures 3 and 4 respectively show the problem and, our solution in schematic form.

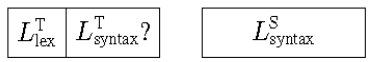


Fig. 3. The Problem.

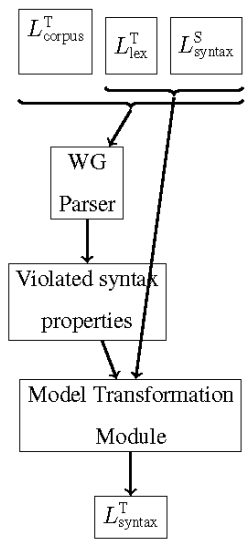


Fig. 4. The Solution.

An Example

Let $L^S = English$ and, L^T = Yorùbá, and, let us assume that English adjectives always precede the noun they modify, while in Yorùbá they are always postpositioned

(an oversimplification, just for illustration purposes). Thus “a red book” is correct English, whereas in Yorùbá we would more readily say “iwe pupa kan” (book, red, a).

If we plug the Yorùbá lexicon and, the English syntax constraints into our WG parser, and, run a representative corpus of (correct) Yorùbá noun phrases by the resulting hybrid parser, the precedence property of L^S_{syntax} will be declared unsatisfied when hitting phrases such as “ìwé pupa kan”. The transformation module can then look at the entire list of unsatisfied constraints, and, produce the missing syntactic component of L^T’s parser by modifying the constraints in L^S_{syntax} so that none are violated by the corpus sentences.

Some of the necessary modifications are easy to identify and, to perform, e.g. for accepting “ìwé pupa kan” we only need to delete the (English) precedence requirement of adjective before noun (denoted by $adj < n$). However, subtler modifications may be in order, after some statistical analysis in a second round of parsing: if in our L^T corpus, which we have assumed representative, *all* adjectives appear after the noun they modify, Yorùbá is sure to include the reverse precedence property of English: $n < adj$. So in this case, not only do we need to delete $adj < n$, but we also need to add $n < adj$.

6.2 Constraint Handling Rule Grammar (CHRG)

Constraint Handling Rule Grammars (CHRG) are a constraint-based grammar formalism added on top of Constraint Handling Rules (CHR) (Frühwirth 1998) similar to the way Definite Clause Grammars are implemented over Prolog. A CHRG works as a bottom-up parser, and, the formalism provides several advantages for natural language analysis. A notable advantage is that the notation supports context-sensitive rules that may consider arbitrary symbols to the left and, right of a sequence (Christiansen 2005).

CHRG (CHR Grammars) strives to incorporate as much as possible of CHR’s expressibility. This includes **propagation**, **simplification**, and, **simpagation** rules and, the capacity for a grammar rule to make considerations for arbitrary grammar symbols that may occur to the left and, right of a sequence that is supposed to match a given nonterminal. Ambiguity is not a problem for CHRG, all different parses are evaluated in parallel in the same constraint store –without backtracking. This results in tagger-like grammar rules that can be used for disambiguating simple and, ambiguous context-free grammar rules, and, provides also a way to handle coordination in natural language. In case of errors, the parser is robust enough to deliver

as its result those subphrases that have been recognized. Since the advent of CHR (Frühwirth 1998) and, of its grammatical counterpart CHRG (Christiansen 2005), constraint-based linguistic formalisms can materialize through fairly direct methodologies.

In a nutshell, CHRG rules rewrite constraints into other constraints, that are subject to possible checks described in a rule's guard and, stated as Prolog calls.

Their general format is:

$$\alpha \ \texttt{-\backslash} \ \beta \ \texttt{/-} \ \gamma \ \texttt{::>} \ G \ \texttt{|} \ \delta \texttt{.}$$

This rule is called a *propagation* rule and, it means, if α occurs before β, and, β before γ, and, G condition is fulfilled, then store δ in the constraint store. The part of the rule preceding the arrow `::>` is called the head, G the guard, and, δ the body; $\alpha, \beta, \gamma, \delta$ are sequences of grammar symbols and, constraints so that β contains at least one grammar symbol, and, δ contains exactly one grammar symbol which is a nonterminal (and, perhaps constraints); α (γ) is called *left (right) context* and, β the *core* of the head; G is a conjunction of built-in constraints as in CHR. If left or, right context is empty, the corresponding marker is left out and, if G is empty (interpreted as **`true`**), the vertical bar is left out. The convention from DCG is adopted that constraints (i.e., extra-grammatical stuff) in head and, body of a rule are enclosed by curly brackets. Special grammatical operators are provided for gaps and, parallel matching (parallel matching is used when it is required that more than one pattern matches the string to be recognized). Gaps and, parallel match are not allowed in rule bodies. A gap in the rule heads is denoted "..." while parallel match is denoted by "$$". Gaps are used to establish references between two long distant elements while a parallel match is useful when more than one pattern is needed to match the string to be recognized (Dahl & Miralles 2012b).

A *simplification (grammar) rule* is similar to a propagation rule except that the arrow is replaced by `<:>` and, the grammar symbols in the head are deleted.

$$\alpha \ \texttt{-\backslash} \ \beta \ \texttt{/-} \ \gamma \ \texttt{<:>} \ G \ \texttt{|} \ \delta \texttt{.}$$

A **simpagation (grammar) rule** is similar to a simplification rule, except that one or, more grammar symbols in its head are prefixed by a ! symbol, which means that such a grammar symbol is not removed. It can also be fully integrated with Prolog and, a single file can conveniently incoporate prolog, CHR and, CHRG codes.

$$!\alpha \ \texttt{-\backslash} \ !\beta \ \texttt{/-} \ !\gamma \ \texttt{<:>} \ G \ \texttt{|} \ \delta \texttt{.}$$

Constraint solving as an implementation paradigm for WG allows us to convey linguistic constraints and, also test them in a modular fashion. The results are also concise.

7 Some Implementation Details

Our CHRG implementation adapts the original implementation (Christiansen 2005) where the appropriate WG constraints are entered in terms of a constraint g/1, whose argument stores each possible grammar property. For instance, our English grammar hybrid parser for noun phrases includes the constraints below and, some others described in section 6.1:

g(np(obligatority([noun, proper-noun, pronoun]))) - at least one noun or, pronoun or proper-noun child
g(np(constituency(determiner))) - children must have their categories in the set
g(np(precedence(determiner, adjective))) - *determiner* children precede *adjective* children
g(np(requirement(noun, determiner))) - if *noun* is a child, then also *determiner* is a child
g(np(dependence(determiner, noun))) - the features of *determiner* and, *noun* are the same

and, so on.

These properties are weeded out upon detection of a violation by CHRG rules that look for them, e.g. an input noun phrase where an adjective precedes a noun may provoke deletion of the constraint g(precedence(noun, adjective)) when the following CHRG rule applies:

```
!word(C2,F1,_):(N1,_N2),...,
!word(C1,F2,_):(_, N4),
{g(np(precedence(C1,C2)))}, {all(A, B)}<:>
{fail(np(precedence(C1,C2)), F1, F2, N1, N4, A, B)}.
```

Fig. 5. Simpagation Rule.

The parser works by storing each Yorùbá word in the lexicon in a CHRG symbol word/3 along with its category and, features (i.e. word(pronoun,[[plural, n/a], personal, both, third-person,[low,mid]], won)). Since the CHRG parse predicate stores

and, abstracts the position of each word in the sentence, this simpagation rule in Figure 5 is triggered when a word of category C2 comes before category C1, given the existence of the grammar constraint that C1 must precede C2 and, the phrase boundaries in which the word is found[2]. Each of the properties dealt with has similar rules associated with it.

When the parse predicate is invoked, the constraints of the source grammar are tested with the input phrases of the target language. Constraints that are violated at least once are output as unsatisfied with the number of times they are unsatisfied and, the features involved. Constraints that are satisfied on the other hand are output as satisfied with the number of times they are satisfied, the features of the words involved and, the phrase boundaries of those two words.

8 Methodology

8.1 Modified Parsing that Calculates both Failure and Success Explicitly

Previous models of WG focused on failure driven parsing, under the assumption that failed properties are usually the complement of those satisfied, so satisfied properties can be derived from the failed ones if needed and, in general, a grammar is induced by repairing failures. However, our more in depth analysis in the context of Yorùbá has uncovered the need for more detail than simply failing or succeeding, as in the case of conditional properties. We therefore now use a success conscious and, failure conscious approach for inducing the grammar of our target language, Yorùbá. Each input phrase of the target language is tested with all relevant constraints for both failure and, success. This makes the model slightly less efficient than if we only were to calculate failed properties, but of course the gain is in accuracy.

Efficiency is still guaranteed by the normal CHRG way of operating: rules will only trigger when relevant, e.g. if a phrase is comprised of only a noun and, an adjective, it will not be tested with, for instance, precedence(pronoun, determiner) or, any other constraint whose categories are different from those of the input phrase. We keep a list of all properties that fail and, another for those that succeed, together with the features of the categories of the words in each input phrase and, the counts of the

[2] Recall that in CHRG syntax the symbols prefixed with exclamation marks are kept, while the ones without are replaced by the body of the rule. The :(A,B) are used to retrive the word boundaries, ... signifies a gap greater or equal to zero, and, all(A,B) retrieves the phrase boundaries.

property (The count of a property is the number of times it occurs, in either the failed list, or, the succeeded list, and, each time a property is added to any of these lists, the count is updated for that property). It is important to state that while other constraints are tested for success and, failure, constituency constraints are tested only for success. This is because we are interested in checking that our target grammar shares similar constituents with our source language and, testing for failure will be irrelevant for these constraints. We also are able to induce constituents present in the target grammar that are not in the source grammar. More of this in section 8.5.

8.2 How we Handle Tones

Yorùbá is a tone language, therefore it was imperative to ensure that tones are properly represented. We adopt a full tone marking approach (as described in section 3.2) so that each word is marked with its tones in order to avoid ambiguities. The tones are also stored as features, so that for instance the word tútù (cold) that has two syllables, one with a high and, the other with a low tone. Each syllable respectively is stored as "High-low" in the features of the word. For instance:

[tútù]::> $word(adjective, [[neutral, n/a], descriptive, both, [High-low]]$, tútù).

In the example, the category of the word is adjective, the features of the word are enclosed in square brackets and, the word is tútù. These tones are used to infer conditional properties in the second phase of parsing. So that if, for instance, a property is found to succeed only if it is around words with certain tones, such property will be said to be conditional. The conditions of such a property will be the tones which form a unique pattern in the success of the property. It is important to note that the tones are not used in isolation of other features and, a property will be said to be conditioned on tones alone, if, tones are the only patterns responsible for succeess or, failure. The addition of tones makes our approach extensible to many other languages including tone languages. In the case of non tone languages, the tones will not be necessary and, can be marked with n/a or may be replaced with any other traits useful for the language.

8.3 Glossing of Phrases

The system provides a record of the English gloss of every word in the target phrases and, this record is consulted during parsing to produce the English equivalent for each word in the input Yorùbá phrase. The gloss is rendered below every Yorùbá phrase, and, the free translation in English comes immediately after. As in:

< 0 > àwọn < 1 > ọmọ < 2 > púpọ̀ < 3 >

```
they     child     plenty
plenty   children
```

Fig. 6. Phrase Transliterations.

8.4 How Conditional Properties are Induced

The original Property-based model, as we have seen, succeeded in detecting and, signalling mistakes in the sentence, without blocking the analysis. In this first model, which is parsing-oriented, incorrect sentences could be "accepted" through declaring some constraints as relaxable. For instance, while from the context-free grammar rules shown in section 5, we wouldn't be able to parse "the the book" (a common mistake from cutting and, pasting in word processors), but this is possible, if we relax the uniqueness of determiner constraint in the constraint-based formulation.

Relaxation can be made conditional (e.g. a head noun's requirement for a determiner can be made relaxable in case the head noun is generic and, in plural form, as in "Lions sleep at night"). The failure of relaxable constraints is signalled in the output, but does not block the entire sentence's analysis. Implementations not including constraint relaxation capabilities implicitly consider all properties as relaxable. And in fact, when exploiting constraint-based parsing for grammar transformation rather than for parsing, this is exactly what we need, since in an unknown language any constraint may need to be "relaxed" and, even if needed, corrected.

For that reason, we have considered it more appropriate, in the context of grammar induction, to state "exceptions" as part of a property itself, rather than separately in a relaxation definition. Thus we have created conditional properties, e.g. "conditional precedence(pronoun, noun)" which expresses that a pronoun preceeding a noun is conditional. We give the condition which says: "if the pronoun is plural and it is a personal pronoun and, can be used in place of both animate and, inanimate nouns". This means that a pronoun preceeds a noun only if the condition holds and, the opposite ordering, if otherwise. We define a property as conditional if it happens that there occurs a pattern for which the property fails for some phrases and, succeeds for others. All constraints that fail and, succeed are tested in this phase. We find these properties by searching for features which are unique to the failure and, or, success of these constraints.The features which are responsible for this difference in behaviour form the condition under which such property succeeds or, fails.

Let us note that requirement(noun, determiner) is in fact a conditional property of English and, the condition under which this requirement property will not hold is if the noun is generic and, plural. So that while count and, or, non generic form of a noun require a determiner (e.g “The boy” and, “The boys”), a plural generic noun does not require a determiner (e.g Chickens lay eggs).

8.5 Failed and Satisfied Properties

For each phrase parsed, the failed and, or, satisfied properties are explicitly output. This enables us to identify which specific properties are satisfied for each phrase and, those which are unsatisfied. We also explicitly retrieve the word boundaries of every word analysed during parsing as well as the phrase boundaries where the words are found. This is basically to eliminate ambiguity that can occur in a situation where a single property fails or, succeeds more than once in a phrase, or, if the model is extended to include other phrases, for example a verb phrase, because the model for now is limited to noun phrases. The boundaries before the semicolon represent the pair of word boundaries while the second boundary after the semicolon represents the phrasal bondaries.

< 0 > àwọn < 1 > ọmọ < 2 > púpọ̀ < 3 >

```
they     child    plenty
plenty  children

Succeeded Property: np(obligatority([noun,proper-noun,pronoun]))1-2; 0-3

Succeeded Property: np(obligatority([noun,proper-noun,pronoun]))0-1; 0-3

Succeeded Property: np(constituency(pronoun))0-1; 0-3

Succeeded Property: np(constituency(adjective))2-3; 0-3

Succeeded Property: np(constituency(noun))1-2; 0-3

Succeeded Property: np(precedence(pronoun,noun))0-2; 0-3

Succeeded Property: np(precedence(pronoun,adjective))0-3; 0-3

Failed Property: np(precedence(adjective,noun))1-3; 0-3

Failed Property: np(requirement(noun,determiner))1-2; 0-3
```

Fig. 7. Parse Results of a Sample Phrase.

In Figure 7, the parse results show that obligatority succeeds twice because there's a noun and, a pronoun in the phrase, three constituency rules succeed because there are three different categories and, two precedence rules succeed. The failed requirement and, precedence properties are also printed.

8.6 Inducing Properties not Present in the Source Language

We also have an additional functionality that finds all precedence and, constituency constraints that are absent in our source but present in our target language.

```
prec(precedence(X,Y)):-
english_properties(List),
not(member(g(np(precedence(X,Y))), List)),
not(member(g(np(precedence(Y,X))), List)),
X \= Y.

word(C1,F1,_):(N1,_),...,word(C2,F2,_):(_,N2),
{others}, {all(A, B)}::> prec(precedence(C1,C2))
|{succeed(np(precedence(C1,C2)),F1, F2, N1, N2, A, B)}.
```

Fig. 8. Inducing Precedence Properties Present in Target Phrase but Absent in Source Grammar.

Figure 8 shows that during parsing, every pair of words with categories `C1` and, `C2`, features `F1` and, `F2` and, word boundaries `N1` and, `N2` are tested for precedence if and only if, the precedence rule of those words' categories doesn't already exist in any precedence rule of the property grammar of English. We define a predicate `prec` which checks that the proposed precedence rule and, its converse is not a member of the property grammar of English while we use `others`, a CHRG constraint to trigger this precedence rule. We do this in order to ensure that we do not create redundancies in a bid to infer precedence rules that occur only in the target language.

The model has the capacity to test the precedence constraints by checking if such a constraint succeeds for all relevant phrases and, if the converse succeeds for any input phrase. If we find that the converse does not succeed in any of the input phrases, this constraint is induced as a property of the target language, else we search for a unique feature to ascertain if it should be induced as a conditional feature. In the case of constituency, like we explained in section 8.1, we are only interested in finding the constituent and, not the number of times it occurs or, otherwise. If a constituent not present in the source grammar is found at least once, in the input phrase of the

target language, then the constituent is induced as part of the grammar of the target language.

8.7 Parsing each Property in our English Subset

Parsing with Dependence Constraints

Due to the structure of Yorùbá grammar with regards to gender and, number, detailed in section 3.5, we initialize all words with neuter gender and, dual number, save a few words which have their number inherent in their meaning and, function (words such as ọ̀pọ̀lọpọ̀ (numerous), àwọn (third person plural). We also implement a predicate that creates a variant for each number or, gender tag based on the environment in which the word is found. This successfully handles the syntactic process of defining number and, gender in Yorùbá. The dependency rule of English is then tested for satisfaction and, otherwise.

This parsing strategy clearly accomodates the existence of plurality in Yorùbá without a previous knowledge of the peculiarities of Yorùbá's syntactic strategies for number, as well as gender which is in reference to English.

However, though plurality and, gender are present in Yorùbá grammar and, are seen to succeed in parse results, the presence of failed dependence relations is further analyzed to determine if the dependency constraints should be a conditional property in Yorùbá. Results show that there is no pattern under which failure or, success occurs. This lays credence to research findings by linguists who have studied the Yorùbá grammar (Awóbùlúyí 1978; Bámgbósé 1966; Ajiboye 2006).

Parsing with the Obligatority Constraint

Previous models of Womb Grammars failed if a noun was absent; the obligatority constraint in our version of WG only fails if a phrase doesn't have a noun or, pronoun or, proper noun constituent as described in section 8.4. Obligatority succeeds if the converse occurs. This idea eliminates the possibility of obligatority failing because it contains a personal noun which is a valid obligatory category.

```
word(C1,F1,_):(N1,N2),{g(np(obligatority(C)))},
{all(A,B)}::> member(C1,C)|
{succeed(np(obligatority(C)),C1,F1,N1,N2,A,B)}.
```

Fig. 9. Rule for Succeeding Obligatority.

In Figure 9, we declare a list of categories `C` which contains a noun, proper-noun and, pronoun. The obligatority constraint is satisfied if `C1` is a member of `C`.

This approach enables us to check if pronouns and, proper nouns in Yorùbá perform the same head function as they do in English. Our results indicate that proper nouns and, pronouns can indeed function as the head of a phrase as in English.

Parsing with Precedence Constraints

In parsing a precedence constraint, we use the precedence rules of English. A precedence rule is satisfied if the target phrase has exactly the same ordering as English. For instance in Yorùbá, every occurrence of `precedence(det,n)`, `precedence(det,adj)` fails while `precedence(adj,n)` fails in certain input phrases and, succeeds in other input phrases (details of this are given in Appendix 1). This gives us adequate information to make a valid conclusion that in Yorùbá determiners always come after nouns and, after adjectives. However, with the presence of both failed and, successful instances of adjectives and, noun orderings, and, a pattern of features responsible, we can only make conclusions that Yorùbá nouns and, adjectives have different orderings. The system materializes this in our output as:

```
-conditional precedence(adjective,noun):-
adjective preceeds noun if adjective is plural,
quantity-uncountable, can be used for both animate and inanimate
nouns
```

Fig. 10. A Conditional Precedence Property.

Our results in Figure 10 reiterate linguistic findings by Yorùbá grammarians (Awóbùlúyí 1978; Bámgbósé 1966).

```
word(C1,F1,_):(N1, _),...,word(C2,F2,_):(_, N4),
{g(np(precedence(C1,C2)))}, {all(A, B)}::>
{succeed(np(precedence(C1,C2)), F1, F2, N1, N4, A, B)}.
```

Fig. 11. Precedence Rule for Checking Success.

In Figure 11 we show from our code that precedence constraints are satisified if and, only if a there exists `precedence(C1,C2)` rule in English, and, an input phrase where `C1` preceeds `C2`. The categories `C1`, `C2` and, the features `F1`, `F2` of the words are stored in a list if the precedence rule is satisfied. We also retrieve the word boundaries and, phrase boundaries as described in section 8.5.

Parsing with Constituency Constraints

A success driven approach is adapted for parsing constituency. This is because we are only interested in the presence of the constituent and, not the number of times it occurs in parsed phrases. Previous Womb Grammar formalisms didn't test for constituents and, this resulted in adding constituents that are absent in the target grammar to the target grammar simply because they exist in the source grammar. We also have a function that can induce constituents that are in the target language but absent in the source, more of this in section 8.6.

```
word(C1,F1,_):(N1, N2),{g(np(constituency(C1)))},
{all(A, B)}::>
{succeed(np(constituency(C1)), F1, N1, N2, A, B)}.
```

Fig. 12. The Constituency Rule.

In Figure 12, every time a word is found, the category of the word is tested with the constituency rule of English. If the category of the word is found in the English properties, then it is satisfied and, added to a succeeded properties list and, subsequently induced. We also store the word boundaries and, phrase boundaries as in section 8.5

Parsing with Requirement Constraints

We test requirement by checking the constituent that co-occur with another constituent in a phrase.

```
check_waits \ wait(Prop, F1, F2, N1, N2, A, B), g(Prop) <=>
fail(Prop, F1, F2, N1, N2, A, B).
check_waits <=> true.

word(Required,F1,_):(N1,N2),
{g(np(requirement(Required,Requiring)))},{all(A, B)} ::>
{wait(np(requirement(Required,Requiring)), F1, [], N1, N2, A, B)}.
{wait(np(requirement(_,Requiring)),_,_,_,_,_,_)},
!word(Requiring,_,_)<:>true.
```

Fig. 13. Rules to Test Requirement Failure.

In Figure 13, we test whether requirement fails by checking that as daughter of the same mother a `Requiring` category doesn't occur with a `Required` category. We achieve this by saving the constraint `wait/7` everytime we find a word with a `Requiring` category and, there exists a property that says `Requiring` category requires a `Required` category. We continue to `wait` till the entire phrase is parsed, if peradventure `Required` will be a category in the phrase which should not necessitate a failure. If no `Required` is found within that phrase, the category `Requiring` and, its features `F2`, its word boundaries `N1` and, `N2` and, the phrase boundaries `A`, `B` where `Requiring` was found are added to the failed properties list; if `Required` is found, we delete `wait/7`.

```
assign(A,B,C,D,A,B,C,D).

!word(X,FX,_):(N1,_N2),...,!word(Y,FY,_):(_,N4),
{g(np(requirement(C1,C2)))},{all(A,B)} <:>
X = C1, Y = C2 -> assign(X, FX, Y, FY, Requiring, FRg, Required, FRd);
X = C2, Y = C1 -> assign(Y,FY,X,FX,Requiring,FRg,Required,FRd) |
{succeed(np(requirement(Requiring, Required)), FRg, FRd, N1, N4,A,B)}.
```

Fig. 14. The Rule to Test Requirement Success.

In testing for success as in Figure 14, if a daughter `Requiring` of a phrase requires a sibling `Required`, we must check that in that phrase (e.g. np) when `Requiring` appears as a daughter, `Required` is also a daughter of that phrase. In the implementation, we check if a `Requiring` and, `Required` occur in the same phrase. The order of the categories is not important here, as we are simply checking if it occurs anywhere within a phrase. We store the features `F1`, `F2` of the word to induce conditional properties in the second phase of parsing. We also store the word and, phrase boundaries which were retrieved as in section 8.5.

9 Results and Supporting Evidence of Correctness

Our results so far have been consistent with linguistic research of Yorùbá grammar which we prove for a fragment of NPs. We also use phrases generated by our Context Free Grammar (CFG), which we wrote for a fragment of NPs considered as the evaluation data in our WG model and, the induced grammar have shown similarities with the CFG. Details of the induced grammar are given in Appendix 1.

It is important to state that despite equivalences that our induced grammar share with the CFG subset, our induced grammar explicitly encodes more information than the CFG. This is because, context free formalisms are unable to directly accommodate extra information such as features, and, their related linguistic constraints, as their grammar symbols can only be simple identifiers, with no possibility of carrying extra arguments, as logic grammar symbols can, to transmit and, consult such extra information.

We summarize our results into constituency, precedence, requirement, dependency and, obligatority.

1 Constituency: Our constituency results show that in Yorùbá, nouns, pronouns, proper-nouns, adjectives, quantifiers and, determiners are allowable categories in noun phrases. These constituents are featured in noun phrases described by Ogunbowale (1970), Awóbùlúyí (1978) and Bámgbósé (1966) as well as in our CFG of Yorùbá.
2 Precedence: We induce two conditional precedence properties (although conditional precedence(adjective, noun) has two different conditions), and, nine precedence properties. Our conditional precedences imply that there are two orderings, which support different orderings in pronouns and, nouns, and, in adjectives and nouns, in for instance, in rules 10 and 11, and, 14 and 15 of our CFG, and, also in literature (Ogunbowale 1970). However, we do not induce a property for quantifier and, noun. This is because, there exist no known pattern responsible for the difference in order, which is consistent with research claims of Awóbùlúyí (1978).
3 Requirement: We do not induce any requirement between nouns and, determiner. This is because of a lack of pattern in features where the requirement property succeeds and, where it fails. This is consistent with our CFG which has rules where nouns occur with determiners, as well as rules where nouns occur without determiners. This conclusion is also presented in research (Bámgbósé 1966; Awóbùlúyí 1978; Ajiboye 2006).

4 Dependency: Our model also does not induce dependency rules. This is because there was no unique feature present with instances where dependency failed and, when it succeeded. This again is supported by Bámgbósé (1966), but not explicit in our CFG of Yorùbá.
5 Obligatority: Obligatority succeeded in all input phrases, showing that at least one of noun, pronoun and, proper-nouns is a compulsory constituent of Yorùbá NPs. Our CFG also shows these three constituents occur at least once in all rules for NPs.

10 Possible Extensions

We covered a useful subset of Yorùbá noun phrases, but there are some phenomena not covered here. For instance, there are some words that are noun phrases with a determiner interfixed between two occurrences of the same words. For example, ọmọkọ́mọ (child any child), ilékílé (house any house), asọkásọ (dress any dress), ìwàkuwà (behaviour any behaviour), ìgbàkigbà (time any time) etc. In certain contexts, these words have different meaning, as in, ọmọkọ̀mọ (child bad child), ilèkìlè (house bad house), ìwàkuwà (behaviour bad behaviour) etc and, these contexts, determine the categories that can occur around them. It would be interesting to develop a parser that accomodates these features. Extending our program to identify and, parse features such as vowel elision, assimilation of tones, compensatory lengthening and, other phonological processes that have grammatical functions in Yorùbá would also be helpful.

Further interesting ramifications of our work would also include: testing our system for a larger coverage of syntax (we have addressed noun phrases so far), inducing other constraints (besides precedence and, constituency constraints which we induced) present in the source language but not in the target language, testing our model on other tonal-sensitive languages, studying how assimilation influences language structure especially in tone languages, studying whether any further constraints or approach extensions would be needed to accommodate families of languages not easily describable within our approach (e.g. languages which have different categories from the source language, those who have different inflectional paradigms from the source language, those that exhibit constraints not among our allowable set of constraints).

11 Conclusions

We have shown the simplicity with which Womb Grammar automatically induces the grammar of Yorùbá from that of English despite the peculiarities in the grammar of Yorùbá and, English that can make this very difficult. This makes our model very useful in language development and, language documentation.

We have also presented a concrete system, for inducing Yoruba grammar that constitutes a proof of concept for the Womb Grammar model (Dahl & Miralles 2012b). WGallows users to input English grammar description in modular and, declarative fashion, e.g, in terms of the linguistic constraints that relate to the subset of noun phrases we consider (constituency, precedence, dependency, obligatority, requirement). Since such descriptions also stand alone as a purely linguistic model of a language's syntax, our system can cater even for users who are not conversant with coding, such as pure linguists wanting to perform syntactic model transformation experiments aided by computers. Their purely (constraint-based) linguistic descriptions of syntax automatically turn into executable code when appended to our code. We also used a subset of noun phrases which can accept any number of adjectives making our model extensible to induce even finer details such as precedence order of adjectives if we so desire.

As we have seen, our system automatically transforms a user's syntactic description of a source language into that of a target language, i.e Yorùbá, of which only the lexicon and, a set of representative sample phrases are known. While demonstrated specifically for English as source language and, Yorùbá as target language, our implementation can accept any other pair of languages for inducing the syntactic constraints of one from that of the other, as long as their descriptions can be done in terms of the supported constraints.

We have thus adapted a flexible modelling framework solidly grounded in linguistic knowledge representation which, through the magic of constraint solving, turns into an efficiently executable problem solving tool. We maintain the modularity feature of the first implementation in our adaptation, therefore, our model can support revision and, adaptation quite straightforwardly: new constraints between immediate daughters of any phrase can be modularly added and, implemented, without touching the rest of the system.

Our model allows for accommodating solution revisions at execution time by interacting with a linguist whenever constraints can not be induced consistently. Visual representations of the output that would be helpful for linguists in this respect have been explored in Adebara & Dahl (2015).

Appendix 1: Results

Due to space constraints, the failed-properties and, succeeded-properties list as well as the phrases parsed have been removed from this results. We present only a subset of the induced grammar.

```
Induced Property Grammar rules for Yoruba Noun phrases::-
-conditional precedence(pronoun,noun):-
pronoun preceeds noun if  pronoun is  plural, personal,
and can be used for both animate and inanimate nouns,
and third-person.
-conditional precedence(adjective,noun):-
adjective preceeds noun if adjective is plural,
quantity-uncountable,  an can be used for both animate and
inanimate nouns.
-obligatority((noun;proper-noun;pronoun)):-
noun;proper-noun;pronoun are obligatority.
It succeeds in all 47 input phrases.
noun succeeds in 39 relevant phrases;
pronoun succeeds in 4 relevant phrases;
proper-noun succeeds in 4 relevant phrases;
-precedence(quantifier,pronoun):-
quantifier preceeds pronoun in all 4 relevant phrases
-precedence(quantifier,adjective):-
quantifier preceeds adjective in all 3 relevant phrases
-precedence(pronoun,adjective):-
pronoun preceeds adjective in all 3 relevant phrases
-precedence(pronoun,determiner):-
pronoun preceeds determiner in all 7 relevant phrases
```

Acknowledgments

This research was supported by V. Dahl's NSERC Discovery grant 31611024.

References

1. Adebara, I. & Dahl, V. (2015). Domes as a Prodigal Shape in Synthesis-Enhanced Parsers. In Kutz, O., Borgo, S. & Bhatt, M. (eds.), *Proceedings of the Third Interdisciplinary Workshop SHAPES 3.0* (pp. 23–33). Larnaca: CEUR Workshop Proceedings 1616.
2. Adebara, I., Dahl, V. & Tessaris, S. (2015). Completing mixed language grammars through womb grammars plus ontologies. In Loiseau, S., Filipe, J., Duval, B. & van den Herik, J. (eds.), *ICAART 2015. Proceedings of the International Conference on Agents and Artificial Intelligence*, Volume 1 (pp. 292–297). Lisbon: SciTePress.

3. Adebara, I. & Dahl, V. (2015). Shape Analysis as an Aid for Grammar Inductions. In Kutz, O., Borgo, S. & Bhatt, M. (eds.), *Proceedings of the Third Interdisciplinary Workshop SHAPES 3.0* (pp. 55–57). Larnaca: CEUR Workshop Proceedings 1616.
4. Adegbija, E. (2004). Language Policy and Planning in Nigeria. *Current Issues in Language Planning*, 5(3): 181–246.
5. Adeyanju, S., Adegbola, T. & Fakinlede, O. (2015). A Supervised Phrase Selection Strategy for Phonetically Balanced Standard Yorùbá Corpus. In Gelbukh, A. (ed.), *Proceedings of the Computational Linguistics and Intelligent Text Processing 16th International Conference, CICLing 2015* Part II (pp. 565–582). Switzerland: Springer.
6. Ajiboye, O. (2006). Interpreting Yorùbá Bare Nouns as Generic. In *Berkeley Linguistic Society Proceedings of the Annual Meeting*, BLS 31.
7. Awóbùlúyí, O. (1978). *Essentials of Yorùbá Grammar*. Oxford: Oxford University Press.
8. Awoyale, Y. (1988). *Complex predicates and verb serialization*. Lexicon Project Working Papers 28. Cambridge: MIT.
9. Awoyale Y. (2008). *Global Yorùbá Lexical Database v. 1.0 Linguistic Data Consortium*. Philadelphia. Retrieved on 28/10/2015 from Natural Language Processing Lab, School of Computing Science, Simon Fraser University, British Columbia.
10. Bámgbósé, A. (1966). *A Grammar of Yorùbá*. Cambridge: Cambridge University Press.
11. Bámgbósé, A. (1993). Deprived, Endangered and Dying Languages. *Diogenes*, 41(161): 19–25.
12. Blache, P. (2000). Property grammars and the problem of constraint satisfaction. In *Proceedings of ESSLLI 2000 workshop on Linguistic Theory and Grammar Implementation* (pp. 47–56).
13. Blache, P. (2004). Property Grammars: A Fully Constraint-Based Theory. In Christiansen, H., Rossen Skadhauge, P. & Villadsen, J. (eds.), *Constraint Solving and Language Processing* (pp. 1–16), Lecture Notes in Computer Science 3438. Berlin: Springer.
14. Blache, P. & Rauzy, S. (2012). Hybridization and Treebank Enrichment with Constraint-Based Representations. In *Proceedings of LREC-2012* (pp. 6–13). Istanbul.
15. Bowen, T.J. (1858). *Grammar and Dictionary of the Yorùbá Language*. Smithsonian Institution.
16. Brill, E. & Magerman, D., Marcus, M. & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop* (pp. 275–282). Hidden Valley: Addison Wesley Longman.
17. Burkett D.& Klein D. (2008). Two Languages are Better than One (for Syntactic Parsing). In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 877–886). Honolulu, Hawaii: Association for Computational Linguistics
18. Chelliah, S.L & de Reuse, W.J. (2010). *Handbook of Descriptive Linguistic Fieldwork*. Berling: Springer.
19. Chomsky, N. (1957). *Syntactic Structures*. Berlin: Walter de Gruyter.
20. Christiansen, H. (2005). CHR grammars. *Theory and Practice of Logic Programming*, 5(4-5): 467–501.
21. Cohen, S.B. & Smith, N.A. (2010). Covariance in Unsupervised Learning of Probabilistic Grammars. *Journal of Machine Learning Research*, 11: 3017–3051.

22. Dahl, V. & Blache, P. (2004). Directly Executable Constraint Based Grammars. In *Proceedings of the Journees Francophones de Programmation en Logique avec Contraintes* (pp.149–166). Angers.
23. Dahl, V. & Miralles, J. (2012a). Womb grammars: Constraint solving for grammar induction. In Sneyers, J. & Frühwirth, T. (eds.) *Proceedings of the 9th Workshop on Constraint Handling Rules* (pp. 32–40). Technical Report CW 624. Leuven: Department of Computer Science, K.U. Leuven.
24. Dahl, V. & Miralles, E. (2012b). Womb Parsing. In Sneyers, J. & Frühwirth, T. (eds.) *Proceedings of the 9th Workshop on Constraint Handling Rules* (pp. 32–40). Technical Report CW 624. Leuven: Department of Computer Science, K.U. Leuven.
25. Duchier, D., Dao, T.B.H & Parmentier,Y. (2014). Model-Theory and Implementation of Property Grammars with Features. *Journal of Logic and Computation*, 24(2): 491–509.
26. Fabunmi, F. & Salawu, A. (2005). Is Yorùbá an Endangered Language? *Journal of African Studies*, 14(3): 391–408.
27. Fagborun, O. (1994). *The Yorùbá Koine - its history and linguistic innovations*. Munchen: Lincom Europa.
28. Fagborun, J. G. (1989). Some Practical Problems in Yorùbá Orthography. *The Journal of West African Languages*, 19(2): 74–92
29. Frühwirth, T.W. (1998). Theory and practice of constraint handling rules. *The Journal of Logic Programming* 37(1-3), 95–138.
30. Merialdo, B. (1994). Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2): 155–171.
31. Nicolas L., Molinero, M.A., Sagot, B., Trigo, E., de La Clergerie, E., Pardo, M.A., Farré, J. & Miquel Vergés, J. (2009). Towards efficient production of linguistic resources: the Victoria Project. In Angelova, G. & Mitkov, R. (eds.), *Proceedings of the International Conference RANLP-2009* (pp. 318–323). Borovets: Association for Computational Linguistics.
32. Ogunbowale, T. (1970). *Essentials of the Yorùbá language*. London: University of London Press.
33. Olumuyiwa, T. (2013). Yorùbá Writing: Standards and Trends. *Journal of Arts and Humanities*, 2(1): 40–51.
34. Pollard, C. & Sag, I. (1994). *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.
35. Romaine, S. (2002). The Impact of Language Policy on Endangered Languages. *IJMS. International Journal on Multicultural Societies*, 4(2): 194–212.
36. Sun, M. & Bellegarda, J.R. (2011). Improved Pos Tagging For Text-to-Speech Synthesis. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5384–5387). Prague.
37. Trudgill, P. (1986). *Dialects in Contact*. Oxford: Blackwell.
38. Voutilainen, A. (2012). Part-of-Speech Tagging. In Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press
39. Wang, Y. & Li, T. (eds.) (2011). *Knowledge Engineering and Management*. Berlin: Springer.

40. Wurm, S. (1998). Methods of Language Maintenance and Revival, with Selected Cases of Language Endangerment in the World. In Matsumura, K. (ed.), *Studies in Endangered Languages* (pp. 191-211). Tokyo: Hituzi Syobo.

Author's Biodata

Ife Adebara is a research masters student at the School of Computing, Simon Fraser University, British Columbia, Canada under the supervision of Veronica Dahl. Ife's research interests are in computational linguistics, translation, constraint programming, constraint handling rule grammars, inducing the grammar of under-resource languages using womb grammars.

TRIANGLE. *Language, Literature and Computation*
Volume 14, 2016, pp. 37–69
DOI: 10.17345/triangle14.37-69
ISSN: 2013-939X
http://revistes.urv.cat/index.php/triangle

Violencia Lingüística en las Redes Sociales

SUSANA CAMPILLO MUÑOZ*

Universitat Rovira i Virgili, Tarragona, Spain

Abstract. Taking into account that, currently, violence is a very frequent phenomenon, this article analyzes a specific type of violence, namely linguistic violence. Specifically, out analysis deals with two topics: the different degrees of violence in speech acts and the resources and techniques that are used to produce violent utterances. By considering politeness studies focused on impoliteness, the work proposes a categorization of violent speech acts that reflects different levels of violence, understanding, therefore, violence as a gradual and non-categorical concept. In the study, a corpus of interactions extracted from Twitter and Facebook is analyzed. The rise of social networks, their ease of use and their features make of these platforms an excellent database for this type of analysis. The results show a significant presence of high levels of linguistic violence in social networks, being the implicitness a feature that in most cases characterizes these utterances. Those results led to reconsider the common idea that defends that explicitness is necessary in order to label an action as violent, and the outstanding presence of high violence categories. It would be necessary to pay more attention to the violent implicit content of utterances which, as shown by the analysis, is not less violent than the explicit one.

Keywords: Violence, language, impoliteness, social networks.

Resumen. Considerando que la violencia es un fenómeno muy presente en la actualidad, en el presente artículo se expone el estudio cuyo objetivo es analizar un tipo de violencia especí-

* **Author's address**:
Departament de Filologies Romàniques
Universitat Rovira i Virgili
Av. Catalunya 35, 43002 Tarragona, Spain
E-mail susana.campillo@urv.cat

fico: la violencia lingüística. En concreto, se ocupa de los grados de violencia, así como de los recursos y las técnicas que se utilizan para emitir actos de habla violentos. Con un análisis de los estudios cartesiológicos, centrado en la descortesía, se ofrece una propuesta de categorías de violencia y diferentes grados. En el estudio realizado se analiza un corpus de ejemplos extraído de las redes sociales Twitter y Facebook; el auge de estas plataformas, su facilidad de uso y sus características convierten a este tipo de plataformas en un excelente banco de datos para realizar estudios como el que se plantea en el trabajo desarrollado. Los resultados obtenidos tras el análisis del corpus muestran la importante presencia de grados altos de violencia lingüística en las redes sociales y la implicitud, en la mayoría de los casos, que caracteriza a estos enunciados. Cabe, quizá, replantearse la necesaria explicitud de las acciones para que sean consideradas violentas, así como la destacada presencia de las categorías de alta violencia. Es necesario prestar más atención a las acciones implícitas que, como demuestra el análisis, no restan carácter violento al enunciado.

Palabras clave: Violencia, lenguaje, descortesía, redes sociales.

1 Introducción

En la actualidad, la violencia se halla en muchos aspectos cotidianos, tanto globales o colectivos como personales o individuales. En los últimos años, el número de acciones agresivas ha ascendido, como muestran los medios de comunicación que diariamente se hacen eco de noticias sobre casos de violencia.

Como acción del ser humano, la violencia puede llevarse a cabo mediante actos de habla que, generalmente, la sociedad percibe como menos violentos que las agresiones físicas. Creemos que, aunque la presencia de daño es necesaria para hablar de violencia, éste no ha de ser ni inmediato ni altamente perceptible. De hecho, la violencia lingüística, más sutil que la física, pasa por ser en muchos casos implícita.

La descortesía, tratada como violencia, ha sido investigada en el ámbito hispánico en las dos últimas décadas (Haverkate 2012; Brenes Peña 2007-2009; Blas Arroyo 2008-2010; Fuentes 2008-2009-2010; Programa EDICE, etc.). Sin embargo, su presencia en las redes sociales no ha recibido una atención especial.

Internet, la Web 2.0 y las redes sociales se han convertido en un gran corpus para el estudio de la comunicación, sobre todo para análisis que adopten un enfoque sociopragmático o sociolingüístico, así como los usos sociales de las tecnologías (Androutsoupoulos y Beisswenger 2008).

Así, las redes sociales se han convertido, desde la década de los 90, en una fuente de corpus lingüístico para el estudio de la comunicación mediada por ordenador (Pano

Alamán y Moya Muñoz 2016; Pano Alamán y Mancera Rueda 2014b). Además, conforman un contexto en el que "han surgido conductas hasta ahora inéditas" (Gómez Arraigada 2013). Éstas permiten el mayor grado de interactividad en la red (Ciber Sociedad 2009): los usuarios pueden comentar publicaciones ajenas con libertad y sin ninguna restricción: en *Twitter* el *hashtag* permite comentar la intervención de cualquier usuario, y en *Facebook*, el hecho de seguir un grupo o página permite a los internautas comentar otras publicaciones en esa página, aunque el usuario emisor "no se tenga en *amigos*[2]".La inmediatez, ubicuidad y espontaneidad llevan al usuario a una revisión casi constante de estas aplicaciones, lo que propicia un incremento de uso.

El estudio que presentamos pretende analizar la violencia lingüística, estableciendo distintos grados y tipos de violencia; analizando las diferentes maneras de manifestarse la violencia en el lenguaje; y describiendo las estrategias utilizadas para enmascarar mensajes inherentemente violentos.

Para ello, nuestro trabajo se ha dividido en dos partes complementarias: una parte teórica, en la que se ha analizado la bibliografía sobre violencia lingüística y las teorías pragmáticas que sustentan nuestro análisis; y una parte empírica, en la que hemos recogido y analizado un corpus de datos propio procedente de las redes sociales.

Las redes sociales seleccionadas han sido Twitter y Facebook. La recopilación del corpus de ejemplos se ha realizado mediante las aplicaciones *Storify* y *Twitter Archiver*[3], y se han recogido 1270 publicaciones[4] (710 de Twitter y 560 de Facebook). El valor del corpus es cualitativo, no cuantitativo: se han seleccionado los comentarios que se han considerado significativos.

A su vez, se ha analizado la literatura sobre violencia lingüística y se ha propuesto una clasificación propia que, seguidamente, se ha aplicado en el análisis del corpus. No se trata de una clasificación genérica, otras clasificaciones serán igual de válidas.

[2] Twitter y Facebook son las redes sociales más populares: Facebook, 81%; Twitter, 43%, según el estudio del VII Observatorio Redes Sociales (16/12/2015).

[3] *Storify* permite seleccionar tuits y comentarios concretos, ordenarlos y descargar el documento que se ha creado; *Twitter Archiver* posibilita la recogida de tuits y retuits de un hashtag de manera masiva.

[4] Los anexos recogen el corpus de datos analizado.

2 Marco Teórico

2.1 Actos de Habla: Austin y Searle

Austin en *Cómo hacer cosas con palabras* (1955) marca la discrepancia en relación con la definición tradicional de *enunciado*, cuya verdad depende de su correspondencia con la realidad objetiva. La base de la teoría que promulga es la distinción entre enunciados *constatativos* y *realizativos*: la misma oración puede ser empleada de manera realizativa o constatativa, han de adecuarse a los hechos o tener cierta relación con ellos. De entre otras teorías, ha estudiado el discurso indirecto como fenómeno social (Trenc 2014).

Seguidamente, sustituye esta dicotomía por tres fases o actos que conforman la emisión de un enunciado: el acto *locucionario*, el acto *ilocucionario* y el acto *perlocucionario*, que corresponden a la realización de acciones (el acto locucionario posee significado; el acto ilocucionario, fuerza, y el acto perlocucionario, efectos). Austin demuestra que *hablar* siempre es *hacer algo* y que un enunciado puede desempeñar diversas funciones, una de las cuales –no la única– es describir.

Searle (1986), discípulo de Austin, presta atención al significado del acto ilocucionario, fruto de la combinación de convención e intención. Así, la emisión de un enunciado en un contexto determinado es realizar convencionalmente una acción concreta (Ruiz Antón 1998). Searle (1976) defiende que es necesario distinguir entre dos fuerzas ilocucionarias: el acto *primario*, que no es literal, y el acto *secundario*, que es literal. Para distinguirlos, el interlocutor ha de ser consciente de que se trata de un enunciado indirecto y urge una estrategia inferencial, en la que entran en juego la *información compartida* (conocimiento léxico y enciclopédico) y la *racionalidad* (dentro de la cual está la memoria).

Aunque no existen reglas que determinen un acto de habla como indirecto, ciertas formas idiomáticas adquieren usos convencionales y son interpretadas como tal si el receptor posee las mismas claves culturales que el emisor (Searle 1976; Escandell Vidal 1993; Terkourafi 2008; Sperber y Wilson 1986).

2.2 Cálculo del Mensaje Implícito: Teoría de la Relevancia

Siguiendo el modelo inferencial, Sperber y Wilson (1986) defienden que el oyente ha de rellenar el enunciado con información extralingüística para, así, obtener una suposición enriquecida contextualmente que corresponde a una proposición (Portolés 2004).

Con fuerte base en la teoría de Grice (1975), Sperber y Wilson toman la máxima de la relevancia como fundamental en la comunicación: siempre se cumple, pues los humanos actuamos según este principio (Albelda y Barros García 2013). Por ello, cualquier acto ostensivo (entre ellos, la comunicación verbal) conlleva la presunción de relevancia; se vincula con el efecto contextual: cuanto mayor es el efecto contextual, mayor es la relevancia.

El concepto de contexto es una representación del mundo compuesta por supuestos, conjuntos de conceptos estructurados de manera que su orden determina el modo en que han de procesarse. El contenido de un supuesto está determinado por las entradas lógicas de los conceptos que contiene, mientras que el contexto en el que se procesa está determinado por las entradas enciclopédicas de dichos conceptos. De este modo, la comunicación se basa en el deseo de alterar (maximizando o minimizando) el entorno cognitivo del interlocutor, de que sus procesos mentales se vean afectados, y el contexto se concibe como producto de la interpretación (Yus 1997).

Para llevar a cabo un procesamiento eficaz es necesario percibir la intención del emisor mediante la construcción de hipótesis, es decir, el reconocimiento de los supuestos a los que el emisor alude para garantizar que su estímulo es relevante.

Un mensaje verbal determina qué conceptos han de activarse y los reúne de forma lógica para que puedan ser usados como suposición (Yus 1997). Así, la comunicación comprende tres niveles:

1. La descripción lingüística (o entrada lógica)
2. Las representaciones semánticas (o entrada enciclopédica)
3. El esquema de supuestos (o derivaciones implicadas).

Para ello se requieren dos procesos pares: la codificación y descodificación, y la ostensión e inferencia.

Según la teoría de la relevancia, con la comprensión se obtiene inferencialmente una suposición enriquecida contextualmente que se corresponde con una proposición. Sperber y Wilson, pues, aportan una vía teórica que explica la comprensión en ausencia de muchos elementos contextuales físicos, puesto que el contexto surge de la combinación del mecanismo perceptivo con la memoria.

2.3 Cortesía: Lo Socialmente Adecuado

Las teorías cartesiológicas definen la cortesía como una decisión del hablante que depende del contexto: se trata de un mecanismo regulador de las relaciones sociales,

no de la comunicación en sí, y el contexto es crucial para interpretarla (Haverkate 1994).

Desde una perspectiva cognitivista (como la de Sperber y Wilson), la cortesía es convención y tiene un papel importante en los supuestos previos que se tienen del comportamiento social adecuado (Albelda Marco y Barros García 2013): se toma la cortesía como lo esperado en determinado contexto. La cognición humana se sirve de dos módulos cognitivos especializados: el sistema cognitivo inferencial y el sistema cognitivo social; se lleva a cabo una doble actividad cognitiva (Yus 2010). La *manifestación de cortesía*, así, depende del contexto sociocultural (Bravo 2008; Brenes Peña 2009b)

La reiteración de comunicaciones forja una imagen cada vez más nítida de normas, intereses, metas comunes y demás que subyacen a la sociedad: crea simetría psicológica (Haverkate 2012).

Teoría de la Cortesía

Brown y Levinson (1987) postularon la teoría que más se ha empleado en los estudios pragmáticos y sociolingüísticos sobre cortesía, y tiene el propósito de complementar al estudio de Grice (1975) ocupándose de los aspectos sociales de la comunicación (Carrasco Santana 1999: 2).

Brown y Levinson definen la cortesía como la inferencia de los interlocutores acerca de las intenciones del hablante; por ello, la forma de los enunciados influye en cómo será interpretado por el receptor en términos de cortesía.

Los interactantes que toman como modelo son competentes miembros de una sociedad que poseen una *imagen* [traducción de *face*, tomado de Goffman (1967)], la propia imagen pública que todos los miembros quieren mantener. Esta imagen se divide en imagen positiva [o *afiliación* (Bravo 1999)] e imagen negativa [o *autonomía* (Bravo 1999)][5].

Para Brown y Levinson la conversación es inherentemente conflictiva, pues amenaza a las imágenes de los interlocutores: algunos actos, como dar órdenes o prometer,

[5] Este modelo de cortesía aplicado al español será estudiado por Carrasco Santana (1999), Garrido Rodríguez (2006) y Haverkate (1994). Bravo (1999) incorpora el elemento sociocultural a la idea de *imagen*, de modo que la cortesía depende del grupo cultural, del contexto sociocultural en el que se realiza dicha actividad, así como de la modalidad discursiva y situacional del momento (Medina López 2011); la cortesía toma un valor relativo a la cultura.

atacan la imagen negativa; otros, como criticar o insultar, la imagen positiva, aunque cabe la posibilidad de que se ataque a ambas imágenes en un mismo acto comunicativo.

Por ello, defienden el uso de determinadas estrategias lingüísticas que mitiguen estos actos amenazantes y proponen una fórmula para evaluar cuál es preferible, en la que se tienen en cuenta la distancia social, la relación de poder y el nivel de imposición permitido culturalmente (W x = D(S, H) + P(H, S) + R).

Con todo, destacan que no todos los hablantes escogerán la misma estrategia en las mismas condiciones, sino que es necesario sopesar cuándo las consecuencias serán menores y qué estrategia dará mejor resultado.

De este modo, dividen los actos amenazantes según si emplean estrategia y según si dicha estrategia es convencionalmente cortés o el enunciado es ambiguo:

- Actos directos [cumplimiento del Principio de Cooperacion de Grice (1975)].
- Actos con estrategia indirecta y abierta[6] (adición de elementos para mostrar que no hay intención de dañar).
- Actos con estrategia encubierta (ambigüedad, se ofrece libertad de interpretación).
- Evitar el acto (desistimiento en la emisión del acto).

Brown y Levinson defienden que el uso del discurso indirecto es la estrategia que asegura la cortesía de cualquier acto, afirmación por la que han sido criticados (Terkourafi 2008; Carrasco Santana 1999; Haverkate 1994; Culpeper 1996). Se trata de la concepción tradicional de la cortesía como expresión de la distancia y estructura social, desde una perspectiva anglosajona.

Teoría de la Descortesía

Considerada, en oposición a la cortesía, como una agresión o amenaza intencional a la imagen del oyente (Albelda y Barros García 2013), la descortesía es un modo de comportamiento con fines estratégicos (Cepeda 2007).

Culpeper (1996) ofrece la contraparte del modelo de Brown y Levinson (1987), siguiendo firmemente los parámetros que establecen (Kaul de Marlangeon 2012). Así, enumera un conjunto de *superestrategias*, en oposición a las estrategias corteses:

[6] En este bloque distinguen entre cortesía positiva, si está orientada hacia la imagen positiva del oyente, y cortesía negativa, si está orientada hacia la imagen negativa del oyente.

- Descortesía directa[7] (ataque claro y conciso).
- Falsa cortesía (empleo de estrategias de cortesía, imposibilidad de atribuir intención al hablante).
- Descortesía encubierta (ataque implícito, el hablante no se muestra responsable del acto) (Culpeper 2016).
- Acortesía (ausencia de cortesía donde se requiere) (Culpeper 2016).

Además, Culpeper (1996) diferencia entre actos directos y actos indirectos:

1. Los actos directos incluyen técnicas que emplean términos con significados propiamente descorteses, que están convencionalizadas en un contexto específico y con efectos perlocutivos concretos.
2. Los actos indirectos se obtienen por derivación inferencial, mediante el contexto o la forma.

Con esta teoría, pues, se defiende que los actos indirectos pueden incrementar la descortesía, pues se fuerza al oyente a un coste de procesamiento extra.

Seguidamente, Culpeper (2011) determina la descortesía convencional y la divide en tres actos comunicativos, del más semántico al más contextual, por el tipo de inferencia que realiza el oyente:

- Derivados por la forma (según el contenido semántico).
- Derivados por convención (según el contexto, parte del enunciado no está acorde con el resto).
- Derivados por el contexto (mediante inferencia, según el conocimiento compartido).

Culpeper explica que las inferencias derivadas del contexto son posibles gracias a su inserción en determinados contextos, que determina su significado y función: las implicaturas que se obtienen son resultado de la unión del acto verbal y el contexto en el que se realizan, más que al propio acto en sí.

Finalmente, defiende que es posible la cancelación de la implicatura –tanto derivada por convención como derivada por el contexto–, ya que está en relación con lo mutuamente manifiesto (Bell 1997), aunque revela la distancia contextual entre hablantes.

[7] Culpeper (1996) también distingue entre descortesía positiva y descortesía negativa, siguiendo el modelo de Brown y Levinson (1987).

2.4 Entre la Descortesía y la Agresión

Tradicionalmente la descortesía se ha vinculado a la violencia, y la agresión, al desequilibrio de la armonía social.

Bousfield y Locher (2008) clasifican la descortesía como una subcategoría de la *agresión verbal* en la que el enunciado es motivado por un *sentimiento personal de rencor* ("personal sense of spite"), en la que la intención no es expresada de manera clara.

Proponen establecer una *escala intencional* con la que graduar la *intención de atacar* que permitiría la distinción y relación entre la agresión verbal, la descortesía y la zona intermedia. Con todo, Bousfield y Locher (2008) aceptan la descortesía como subcategoría de la agresión verbal.

Terkourafi (2008), por su lado, afirma que tanto la cortesía como la descortesía no se realizan de manera directa porque con ellas se *evalúa* el comportamiento del hablante, y lo que determina si un acto fortalece o amenaza la imagen son sus efectos perlocutivos. Distingue, también, entre grosería implícita (convencional en relación al contexto), grosería explícita (no convencionalmente relacionada con el contexto) y descortesía (no convencionalmente relacionada con el contexto, ataca la imagen, pero no es posible atribuirle la intención al hablante). De este modo, lo que diferencia la descortesía de la grosería es que esta última sí que inherentemente daña la imagen.

Sin embargo, Terkourafi (2008) afirma que, aunque las expresiones queden fijadas convencionalmente y se les atribuya determinado efecto perlocutivo, no siempre producirán ese efecto, sino que dependerá del hablante; todo ello hace menos fija la descortesía.

2.5 Violencia y Lenguaje

Vista como fuerza primaria y elemental en el ser humano (Bornhauser 2012), la violencia se entiende como un acto que atenta contra la libertad interior y exterior del individuo (Fuentes Rodríguez y Alcaide Lara 2009). Es un elemento propio de la cultura y de las relaciones sociales [Kolakowski 1999 (citado por Colín Rodea 2011)].

La relación que la violencia mantiene con el lenguaje es destacada: conlleva la ejecución de un acto ilocutivo destinado a dar la imagen del interlocutor (Fuentes Rodríguez y Alcaide Lara 2008). La hostilidad de cualquier tipo de acto de habla

depende del contexto, del *contenido* y de la interpretación; es un tipo de comportamiento.

No obstante, si se trata la violencia como descortesía extrema, los actos verbales solo se clasifican como violentos si son explícitos, aunque autores como Brenes Peña (2009a) defienden que la explicitud no garantiza descortesía.

Las teorías tradicionales han considerado los actos descorteses como perlocutivos, pero, si se afirma que la descortesía, así como la violencia, es una estrategia para alcanzar un fin, necesariamente ha de considerarse la intencionalidad del emisor de realizar dicho acto.

Por otro lado, los principios que regulan la descortesía están en evolución constante (Fuentes Rodríguez y Alcaide Lara 2009), y el resultado es la variabilidad de parámetros en la categorización de la descortesía (como el cultural o el personal) (Balandrón Pazos 2004, citado por Brenes Peña 2009a).

Los estudios sobre violencia, descortesía y lenguaje son heterogéneos: según el autor, la agresión se distingue de la violencia, pero no la violencia de la descortesía ni la descortesía de la agresión (Fuentes Rodríguez y Alcaide Lara 2009; Blas Arroyo 2010; Fuentes Rodríguez 2010; Mancera Rueda 2015; Brenes Peña 2009b). En general, no quedan establecidos unos límites claros: se puede tomar en consideración la intención, el fin (Balandrón Pazos 2004) o el grado de lesión (Kaul de Marlangeon 2008), por ejemplo.

2.6 Redes Sociales

Por su particularidad, las redes sociales conforman un contexto distinto al de la comunicación oral: el canal, la distancia temporal entre hablantes y la información contextual son distintos (Díaz Pérez 2014).

En Internet y, sobre todo, en las redes sociales encontramos un tipo de comunicación que no puede clasificarse ni de oral ni de escrita; se trata de un híbrido de ambas, –un *texto oralizado*, en términos de Yus (2010)– que Crystal (2002: 62) denomina *Netspeak*, “lenguaje escrito que se ha estirado en dirección al habla”. Es la gente quien influye de forma creativa en este lenguaje virtual, renegocian los géneros a través de la interacción social (Pano Alamán y Mancera Rueda 2014a; Iglesias Moreno 2012).

Además, la temporalidad, el ritmo, la no respuesta inmediata y el paralenguaje difieren de los de la comunicación cara a cara (Lambert Graham 2008). Aun así, el

contexto virtual ofrece unas variables de interpretación distintas, pero igual de fructíferas, tales como emoticonos o recursos tipográficos.

El registro más frecuente es el coloquial y, pese a ser tecleado, no se presta la atención ni se realiza la revisión exhaustiva propia del texto escrito. Se trata de un texto escrito "sin filtro", "improvisado", pero que, a su vez, goza de una planificación previa: aparentemente es natural, pero el usuario, antes de emitir –o publicar– el enunciado ha podido leerlo y revisarlo (Mancera Rueda 2015).

¿Descortesía en Internet?

Las características del mundo virtual crean un contexto en el que la no proximidad personal lleva a la desinhibición y a la liberación de la responsabilidad de la ofensa o el ataque. No es de extrañar que abunde la descortesía, pero no se ha de relacionar con la violencia. El fin es, en gran medida, lúdico y de afiliación (Díaz Pérez 2014), de afianzamiento del vínculo de los receptores, que queda marcado, sobre todo, en el *Retuit* o el *Me gusta*.

La temática del momento –derivada, sobre todo, de noticias que causan agitación– es otro de los factores que encaminan a los usuarios a la descortesía y a la agresión: es común hallarla en temas políticos e ideológicos (Vivas Márquez y Ridao Rodrigo 2015).

Así, parece claro que Facebook y Twitter son espacios favorables tanto para la descortesía –puesto que suponen una libertad expresiva y un mecanismo de desfogue emocional– como para la agresión –ya que se trata de un mundo ficticio y virtual que parece despersonalizar cualquier acción–.

3 Violencia Lingüística: Propuesta de Clasificación

Debido a la variedad de criterios considerados en las diferentes clasificaciones de violencia lingüística, hemos creído oportuno establecer una propuesta de clasificación para realizar el análisis (Figura 1).

Los dos extremos que conforman el continuum son la grosería y la agresión. Ambas se consideran actos violentos en diferente grado: la *grosería* ataca únicamente la imagen del hablante; la *agresión*, la imagen del hablante y la del oyente. De esta manera, la agresión se concibe como el grado máximo de violencia y la grosería,

Fig. 1. Continuum de Violencia.

como el mínimo, formando así un continuum con diferentes niveles/grados entre ambos: desaprobación, ridiculización, rechazo y discriminación, insulto y amenaza.

Los parámetros que distinguen grosería y agresión son la *variabilidad* y la *generalidad*: la *grosería* varía entre culturas, contextos e interlocutores –se caracteriza por la subjetividad (Lambert Graham 2008)–, mientras que la *agresión* es tomada generalmente como tal en cualquier situación y por cualquier hablante (salvo alguna excepción). La grosería supone la transgresión de las normas de conducta culturalmente establecidas y esperadas por los hablantes en cualquier situación. La agresión es la exteriorización de la agresividad como ataque al interlocutor

Ambos extremos son intencionales, el contexto de las redes sociales favorece la percepción intencional (lenguaje escrito y leído simultáneamente), aunque, respecto a la grosería, la permisibilidad de la mala educación varía entre culturas y clases sociales.

Desde una perspectiva lingüística, se tiene en cuenta que los actos violentos, como actos de habla, se pueden realizar de forma explícita o implícita, distinguiendo, así, tres niveles en los que hallar el contenido violento: el contenido violento inferido *por la forma*, que presenta el mayor grado de explicitud; el contenido inferido *por el contexto*, que representa un grado intermedio; y, por último, la violencia inferida por convención (véase Culpeper 2011).

Los recursos de las diversas categorías del continuum son varios y, en muchas ocasiones, compartidos en los actos lingüísticos violentos. Su uso es posible gracias al empleo las implicaturas conversacionales, las implicaturas convencionales y el conocimiento compartido, tanto cultural como tecnológico.

Así, en una gradación de lo directo a lo indirecto, estos mecanismos son:

1. Comparaciones
2. Exageraciones
3. Ironías
4. Sarcasmo
5. Interrogaciones y exclamaciones retóricas

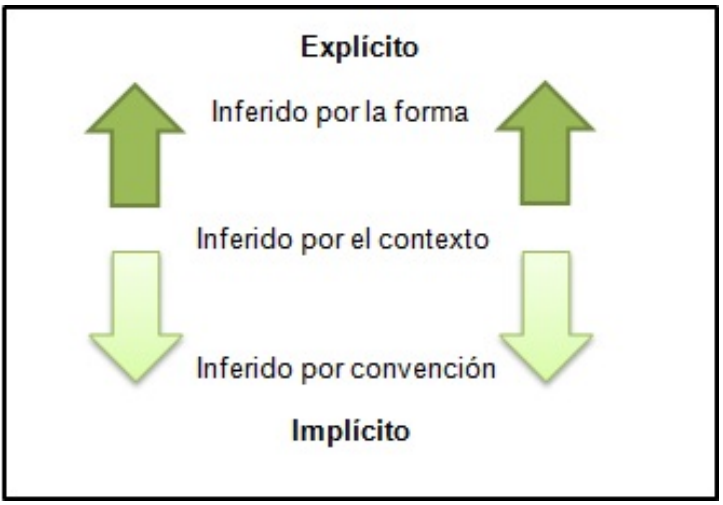


Fig. 2. Gradación Explícito e Implícito.

6. Voces de eco o descripciones.

4 Categorías del Continuum de Violencia: Análisis

Teniendo en cuenta la clasificación de violencia lingüística propuesta, hemos realizado un análisis del corpus recopilado.

4.1 Groserías

El uso del lenguaje soez y de las palabras malsonantes, pese a ser visto en varias culturas como descortés (Culpeper 2011; Terkourafi 2008), se ha convertido en un hábito lingüístico, una especificidad idiomática de la cultura española como símbolo de libertad de expresión (Rius 2012). Por ello, su empleo ataca a la imagen del emisor, a quien se le atribuye mala educación. Su normalización junto al distanciamiento de internet facilita la interpretación del lenguaje soez como no amenazante para la imagen del oyente.

La grosería se asocia a la defecación, la sexualidad y la digestión, considerándose como "sinónimo de vulgaridad" (Colín Rodea 2011: 12). El Diccionario de

la Lengua Española (2016) y el Diccionario CLAVE, por su lado, definen grosería como "descortesía, falta grande de atención y respeto" y "rusticidad, ignorancia".

La categoría de *groserías* incluye, pues, palabras malsonantes (*coño, cojones, mierda,...*), de las cuales muchas han perdido la correlación con el ámbito escatológico o sexual, fijándose como "tacos":

(1) Todos revueltos en la mierda.

También incluye palabras tabúes u obscenas en determinada cultura (*hostia, puta,...*)[8]:

(2) Hay q joderse el amiguismo q sale por cualquier sitio.

Las **expresiones relacionadas con el** *asco* son el paso del lenguaje soez a la desaprobación o desacuerdo, porque el efecto de "provocar asco" se relaciona con el ámbito escatológico y digestivo:

(3) Las joyitas de #microhembrismosson grandes eh, y con joyitas quiero decir ganas de vomitar.

Aunque la grosería sea inherentemente marcada, se puede encontrar, con los recursos tipográficos al alcance, enunciados groseros indirectos en los que el **término soez aparece atenuado** –mediante juegos fonéticos o el empleo de símbolos tipográficos– como también enunciados indirectos simples, que se sirven de implicaturas derivadas del contexto:

(4) Eso de que # LosHomosexualesPeganSIDA es una # ¡ $ % :)
(5) Mi perro, paseando, me acaba de dar su opinión sobre el #CaraACaraL6. La he recogido con la bolsa y tirado a la papelera.

Junto a los insultos, el lenguaje soez es una fuente de información sobre las convenciones y normas lingüísticas impuestas por los hablantes (Bernal Linnersand 2007).

[8] En nuestro corpus, los tacos "expresan emociones tales como la ira, la excitación y también actúan como refuerzo" (Rundblom 2013); estos términos muchas veces se complementan con el uso de otras estrategias como refuerzos expresivos.

4.2 Desaprobación

La muestra de discrepancia con otros usuarios conlleva un grado de ataque a la imagen del receptor, pero la distancia entre usuarios y la libertad expresiva de internet atenúan el posible ataque. Aun así, la confrontación de ideas y opiniones fomenta las relaciones sociales y crea lazos positivos (Bravo 1999), y la cultura española [respecto a los suecos, por ejemplo (Fant 1989)] tolera grados elevados de desacuerdo.

En el nivel de lo explícito, esta categoría incluye fórmulas (*no apruebo*), expresiones del efecto que produce determinado acto en el receptor (*me sigue encendiendo*) o juicios directos al usuario; se emplean implicaturas conversacionales y, por tanto, se obtiene el contenido mediante la derivación del contexto:

(6) Vergüenza ajena de oír hablar al marido de la #pedroche #estoeslahostia solo sabe decir eso
(7) Aparte de que lo veo una irresponsabilidad enorme . . . Lo mejor de toso los twits es "soy torero por la gracia de Dios" Dudo mucho que a Dios le guste la manera de torturar toros que tienes

El desacuerdo se puede expresar indirectamente, como se ha comentado, mediante implicaturas contextuales o conversacionales. En éstas se pueden observar expresiones populares, reacciones del usuario frente a la acción y generalizaciones:

(8) Con la que cae y . . . #CompiYogui
(9) Definitivamente, he perdido la esperanza en la humanidad #LosHomosexualesPeganSIDA

El recurso más empleado en las expresiones de desacuerdo es el de los **enunciados de eco** (derivados del contexto), interpretaciones de un pensamiento de otra persona que informan de que el emisor tiene en mente lo que tal persona dijo y una determinada actitud hacia ello (Sperber y Wilson 1986):

(10) Gensanta! Escuchar a Rajoy decirle a Sánchez que tiene un problema en los ojos. #CaraACaraL6

Las **ironías** se incluyen en los enunciados de eco, aunque la actitud expresada en un enunciado irónico es siempre de rechazo o desaprobación (Sperber y Wilson 1986); no es necesario elementos explícitos que marquen la actitud hacia el pensamiento descrito:

(11) #LosHomosexualesPeganSIDA y catarros. Y ojo cuidado con los catarros.

Las **comparaciones** y las *interrogaciones retóricas* también son numerosas en nuestro corpus. Las primeras, equiparan la situación con otra con la que comparte el elemento que se desea criticar; en las interrogaciones retóricas se formula una pregunta cuya respuesta sería de negación, mostrando sorpresa y discrepancia:

(12) #micromachismos "Chicas no dejéis que los novios controlen vuestro teléfono." Soraya Sáenz de S. María. ¿La responsabilidad de la mujer.?

En la zona de lo implícito, finalmente, se encuentran las **descripciones**, enunciados que definen el elemento que se critica cuyo desacuerdo se infiere por la perspectiva que toma el emisor en la descripción:

(13) Gente aplaudiendo como una persona agrede por la espalda a otro que no se lo espera. #PuñetazoARajoy

La discrepancia junto al ataque a la imagen explícito se materializa en la **ridiculización**.

4.3 Ridiculizaciones

La ridiculización supone un grado más que las anteriores categorías en el continuum de violencia lingüística, porque implica de forma evidente un ataque a la imagen –positiva–del oyente; la ofensa es segura.

El *ridículo* es definido por el *Diccionario Salamanca de la lengua española* (2006) como "algo que por su rareza o extravagancia mueve o puede mover a risa", marcadamente peyorativo, porque "provoca o puede provocar la burla de los demás".

El uso del humor como herramienta de destrucción de la imagen del interlocutor (Bañón 2010) se ejemplifica en las ridiculizaciones: las **burlas**, asociadas al humor, atacan de forma clara la imagen –pública sobre todo– del interlocutor.

(14) fran Rivera es tauromaquiavelico

Del mismo modo que en las comparaciones, se asocia a la persona con determinada situación cómica o bien resaltando determinado aspecto de manera burlesca, de modo que denotan cualidad contrarias a las que cabría esperar. En muchos casos, la ridiculización se deriva mediante la inferencia de implicaturas contextuales:

(15) Es una intelectual del siglo XXII (si si lo he dicho bien) es una avanzada a su tiempo, una obra maestra, a su lado Mozart y Salieri son meros aprendices Una letra cargada de mensaje, e hace reflexionar ¿De dónde venimos? ¿A dónde vamos? ¿Quiénes somos? ¿Si le saco todo el plástico que lleva en el cuerpo me podré hacer un IronMan de juguete a tamaño real?

En varios casos, los elementos de la comparación dependen de la cultura, de modo que la implicatura se infiere por convención (por ejemplo, el término *colorear* se asocia en la cultura española a la infancia):

(16) Sanchez ya colorea sin salirse del borde. #CaraACaraL6

Del mismo modo que en las expresiones de desacuerdo, en las ridiculizaciones es común también encontrar **interrogaciones retóricas**, pero en estos casos funcionan como mecanismo de comparación con situaciones dispares; en estas interrogaciones retóricas la comparación se hace de manera implícita, mediante la ambigüedad léxica o estructural (*diera la cara*, por ejemplo) o utilizando expresiones que forman parte de la cultura española:

(17) ¿Te gustan los sueldos altos? Pues vas a cobrar #PuñetazoARajoy
(18) ¿Los toreros no tienen madre que les diga que porque #FranRiverase tire por un pozo no tienen que tirarse ellos?

Son destacadas las ridiculizaciones que apelan a elementos de actualidad tecnológica, tales como programas de humor, los ordenadores o los propios "memes"[9]:

(19) Rajoy error code #CaraACaraL6
(20) vamos a darles un aplauso y dejamos que se vayan x'Djajajajajaja
(21) Que no han trucado la foto del puñetazo de Rajoy para q parezca peor, que la han editado en insta y han subido el contraste #punetazoarajoy

Los recursos tipográficos se emplean, sobre todo, como ataque a la ortografía del usuario o para exagerar la pronunciación de forma marcada, en un tono ridículo:

(22) HA ESTADO***

[9] Construcción multimedia que describe situaciones, ideas o pensamientos de forma cómica y burlesca; son características de las redes sociales.

Como muestra del componente cultural de la pronunciación y la entonación, el siguiente ejemplo contiene varias ridiculizaciones con una pronunciación "españolizada" en inglés, que connota ignorancia, o exageradamente inglesa, que connota cursilería o arrogancia:

(23) Con mis zapatos Manolo's Puedo? Puedo?Porfi!!!! Yo también sé decir "Darling osea".
(24) Of coursedear! Lo voy a a apuntar en mi agenda: después de Pilates y del masaje. Qué cool tía!
(25) Al Senado porque yo lo valgo.Rabia, rabiña
(26) Y el Brunch en ese sitio tan divino con esos camareros tan monísimos en Lagasca, OK?
(27) Puedesssss! Churri, ossssseaa, quedamos aána para el branxxx? (brunch) porfiiiiii?
(28) Anota en tu aipaz. Y luego me llamas desde tu aifon seis plas. Sí, sí, plas.
(29) Pero el plas es mucho mejor que el plus, no lo dudes! No olvides que estudié en Quembrich!
(30) Yo también tengo un iphone6. Y en cuanto me entero q algún contacto mio no tiene le bloqueo.

Expresiones como "of course", "Darling osea" o "Qué cool tía!" se asocian a personas pijas o de alta clase social; otras como "rabia, rabiña", "jopelines" o "porfiiiii", niñez; y la pronunciación exagerada como "Oxfor" o "aifon seis plas", cursilería y petulancia. Así, estos casos forman parte de las ridiculizaciones implícitas derivadas por convención.

4.4 Rechazos y Discriminaciones

El rechazo y desprecio al interlocutor supone un ataque directo a su imagen, tanto positiva como negativa. Se alude a la persona de forma marcada, a pesar de que los mensajes en internet están en realidad destinados a la colectividad (Vivas Márquez y Ridao Rodrigo 2015). Es común encontrar en estos enunciados groserías; el uso del lenguaje soez es empleado como recurso expresivo y, por tanto, aumenta el grado de exhortación:

(31) Y ahora ya que se vayan por el mediterráneo a otro país, a tomar por culo de aqui.

En la zona de lo implícito, las comparaciones provocan distintos efectos según el elemento con el que se compare al sujeto –como sucede en la desaprobación y en la ridiculización–; en el siguiente ejemplo, se ridiculiza y menosprecia al usuario:

(32) Pilar Rubio, el interés que me produce tu vida personal es igual al interés que muestro por la reproducción del percebe. #SospechaEH

Las **interrogaciones retóricas** expresan de forma indirecta el descontento del usuario en relación a determinada acción o al sujeto que la realiza, y el **sarcasmo**, "burla sangrienta, ironía mordaz y cruel con que se ofende o maltrata a alguien o algo" (como define el Diccionario de la Lengua Española) expresa, mediante un enunciado en eco, el desacuerdo y rechazo, cuyo fin es inherentemente ofensivo, a diferencia de la ironía.

(33) no sabia que los reyes tienen amigos tan rinbonbantes, por dios hasta cuando
(34) por favor no quiteis por asuntos terroristas #myhyva la juventud iberica que se despierta a la1 para verlo y se vuelve acostar hasta la cena.

Las discriminaciones, las acciones que dan "trato desigual a una persona o colectividad por motivos raciales, religiosos, políticos, de sexo, etc.", como define el Diccionario de la Lengua Española, conllevan un rechazo grupal, que supone un ataque a la imagen positiva y negativa de cada uno de los integrantes del grupo.

En las **discriminaciones implícitas derivadas por convención** destacan las ovaciones a personajes o acontecimientos históricos que fueron contrarios a las personas –o grupos– que se rechazan o discriminan. Los personajes y hechos históricos a los que se alude, además de ser inherentemente discriminatorios, se asocian a actos físicamente violentos, hecho que incrementa la violencia del enunciado.

(35) Gitanos un trabajo que le quedo pendiente al Canciller aleman.
(36) Robespierrevuelveee!

Los enunciados de eco en las discriminaciones están generalmente asociados a determinados grupos sociales, comunes en enunciados xenófobos y machistas. En muchos de los ejemplos en los que se emplea este recurso se altera el sujeto, de modo que, además de discriminar, se ridiculiza al usuario:

(37) mi jefe me dijo el otro día que me afeitase la barba. A ninguna de mis compañeras le ha pasado. #micromachismos

En el caso de la discriminación por sexo, muchos de los enunciados contienen groserías, lo que incrementa el rechazo, denigrando al sexo al que se ataca:

(38) Ojalá darle un pollazo en la frente a cada uno, por la cosa de la #igualdad. Stop #micromachismos.

4.5 Insultos y Amenazas

Los insultos y las amenazas conllevan un ataque directo a la imagen del interlocutor: los insultos atacan a la imagen positiva; las amenazas, a la imagen negativa. Esta es la principal distinción, pues el grado de violencia es el mismo: cerca del máximo.

Los rasgos que caracterizan a los insultos son la oralidad –traducida por escrito con diacríticos, mayúsculas, puntos suspensivos e interjecciones–, la ambigüedad buscada intencionalmente por el hablante, la inestabilidad y sincronismo, y la gradación, sujeta a las acepciones y estructuras en las que se encuentran (Bernal Linnersand 2007).

Componentes de los insultos que son inherentes a la violencia son la salud (véase Campos Santelices 2010) y los derechos humanos. Se atacan las capacidades humanas del individuo (inteligencia, sexualidad, educación), así como aspectos relativos, rasgos que se consideran positivos (bondad, sinceridad, respeto):

(39) SOIS UNA PUTA ESCORIA #MYHYV
(40) eres una troll o eres una auténtica retrasada. No sé porqué, pero me da por lo segundo . . .
(41) Vale, vale, tomamos nota realeza. Realeza o mentireza!!

Las generalizaciones, mediante implicaturas conversacionales, conllevan un distanciamiento del emisor con el enunciado y el insulto, y las acusaciones conllevan la asociación de una etiqueta o característica que, en la mayoría de casos, es un insulto. Se trata de insultos cuya interpretación se deriva del contexto, en los que la etiqueta no marcada hace que el emisor eluda responsabilidad en el acto.

(42) #LosHomosexualesPeganSIDA si tirases una piedra en este hastag encontrarías 20 subnormales por m^2
(43) #LosHomosexualesPeganSIDA ¿Qué narices? La Iglesia Católica tiene demasiado acceso a Twitter!

Entre las técnicas que se emplean para insultar, las que ocupan un mayor número de ejemplos en el corpus analizado son la **ironía**, el **sarcasmo** y las **comparaciones**:

(44) Lo reconozco, no me dice nada nuevo este nuevo debate entre Cánovas y Sagasta #CaraACaraL6
(45) Prefiero ver el bebé en brazos de Fran Rivera toreando, que de tropezón en un gintónic de Rita Barberá
(46) Me encanta el retraso que tenéis. #LosHomosexualesPeganSIDA

También se utilizan chistes, entre los que pueden incluirse las canciones populares:

(47) Ande andeande País de Pandereta Ande andeande Que no tienen JETAAA #CaraACaraL6 #CaraACaraTVE

Los recursos tipográficos (pronunciación exagerada, puntos suspensivos y otros) son un mecanismo de mitigación que "esconden", en lo implícito, el insulto (como se ha visto, también se emplean en el lenguaje soez y las ridiculizaciones):

(48) Pero que se puede esperar de "gente" que ve programas como #Myhyv?
(49) Espero k mañana no kiten mueres y omvres y bisevelsajolinkon los atentaos ke pesaos los lokos estos #myhyv

Los insultos más indirectos son los que apelan al conocimiento compartido (derivados por convención), inferidos, así, mediante implicaturas convencionales y culturales:

(50) Lo dicho, hoy Darwin se tendría que tomar un Valium.
(51) Fran Rivera el que lo lea . . .

La violencia de estos enunciados se percibe solo si se tiene conocimiento de estos aspectos, por tanto, es relativa, mientras que el insulto directo contiene una interpretación peyorativa general (pese a la relatividad cultural).

La percepción de los insultos, con todo, varía en función del contexto; en las redes sociales –en los hashtags y publicaciones de grupos investigados–, debido a la distancia y el desconocimiento de usuarios, el insulto se percibe como un ataque.

Las **amenazas** se definen como "el anuncio al receptor de nuestra voluntad de llevar a cabo o ejecutar en un tiempo futuro más o menos próximo un acto o acción que

supone un daño hacia su persona" (Brenes Peña 2009c). Atacan a la imagen negativa del interlocutor, aunque suelen ir acompañadas de ataques a la imagen positiva, de insultos, y apelan a la violencia física de forma explícita:

> (52) hijos de puta!!!!!!!! y yo para ganar 8 mil euros tengo ke sufrir mobbing en citroen y luego en la mutua por parte de los psicologos y luego en la inspeccion-medika hijos de puta todos eniaisketemrinar todos colgados en la puerta de l sol de una soga y linchamiento publiko e reinita ositayogui!!!!? hijos de putqa ladrones algundiaardereis!!!!!!! y os trituraremos vuesroscrneoskeuaremos de meaderos publikos –un beso compi yogui hay que ser subnormal puta sifilitika de mierda ke no te vea como rajoy por pontevedra o asutriaske vas a chupar igual kerajoy . . . . . .

El emisor, en las amenazas indirectas, elude la responsabilidad del acto y puede que no se dirija directamente al sujeto en cuestión:

> (53) Si queréis acabar con el bipartidismo, pasadme las coordenadas del plató. #CaraACaraL6

Dentro de las amenazas, las maldiciones, deseos de sufrimiento y daño hacia el interlocutor son una forma atenuada de amenaza, en la que el interlocutor no se incluye como actante en la acción violenta –en analogía con las acusaciones y los insultos–. Las maldiciones pueden realizarse mediante el uso del subjuntivo (por ejemplo, con la expresión desiderativa *ojalá*) o bien con el empleo del indicativo (por ejemplo, con la afirmación de que el dolor que se impreca acaecerá):

> (54) No es más tonto por q el día no es más largo. Ojalá se reuna pronto con su dios q menudo padre protector esta echo el memo
> (55) Eres descpreciable @Milaximenez te alegra de todo el mal de las personas ya te llegará el turno #campostelana @salvameoficial

Las amenazas y los insultos representan un grado elevado de violencia, atacan directamente a la imagen positiva y negativa del destinatario.

4.6 Agresión

Tanto la muestra de acuerdo con una acción violenta físicamente como la invitación a que la acción se realice se incluyen en esta categoría, que recoge aquellas expresiones a favor de la agresión física. El ataque a la imagen del interlocutor es evidente.

Se observan ejemplos que aprueban de forma abierta la acción violenta, pero también es posible realizar dicha aprobación de forma indirecta (por ejemplo, con inferencias por el contexto):

(56) A ver, relajaos . . . que ha sido un golpe de nada. Pocos puñetazos recibe el señor ese con todo lo que ha liado! #punetazoarajoy
(57) 1,5mL de "Hostión 300" a mano abierta, vía facial, tres veces al día; desayuno, comida y cena. En casos extremos (como es el caso), doblar la dosis a 3mL, dos tomas en desayuno , dos en la comida y dos en la cena. Repetir el tratamiento durante un año

La diferencia de esta categoría con la de *amenazas* reside en que, en este caso, el emisor no se implica explícitamente en el desarrollo de la acción (con el empleo del subjuntivo, la 3ª persona del singular o la forma impersonal) y el destinatario, en varios casos, tampoco queda definido individualmente:

(58) Entiendo que uno tuvo error de tipografia . . . pero los ultimos 2 de arriba a abajo, izquierda a derecha, habría que golpearlos hasta que les deje de doler . . .

La apelación a noticias de actualidad, por ejemplo, caracteriza a los ejemplos de agresiones implícitas, de modo que, mediante implicaturas convencionales, culturales o de actualidad, es posible comprender el significado violento:

(59) Pena me da que los seguidores de #myhyvno vayan en autocar

Finalmente, en esta categoría es posible, también, hallar ejemplos en los que se emplee la tipografía como recurso para expresar pronunciación marcada, escondiendo en lo implícito el acto violento al que se apela:

(60) #CompiYoguicompi yogui compiyoguicomguiyopicomguillopi con guillopi con guilloti con guilloti na con guillotina

Así las cosas, un acto marcadamente explícito y directo como es la agresión se puede realizar de manera implícita.

5 Consideraciones Finales

5.1 Resultados

El objetivo principal de este trabajo ha sido analizar la violencia lingüística y para ello nos hemos servido de un corpus extraído de las redes sociales. Los objetivos que nos planteábamos eran los siguientes:

- Averiguar cómo se establece la relación entre la violencia y el lenguaje.
- Establecer distintos grados de violencia lingüística.
- Determinar qué características poseen los enunciados violentos.
- Analizar las técnicas empleadas en las redes sociales en la emisión de enunciados violentos
- Indicar los tipos de actos violentos que predominan en las redes

La teoría de los actos de habla afirma que hablar es realizar un tipo concreto de acto, de modo que es posible actuar violentamente al hablar. El hecho de que el corpus propio analizado se haya extraído de las redes sociales (plataformas virtuales en las que se interactúa principalmente con palabras) corrobora la existencia del acto de habla violento.

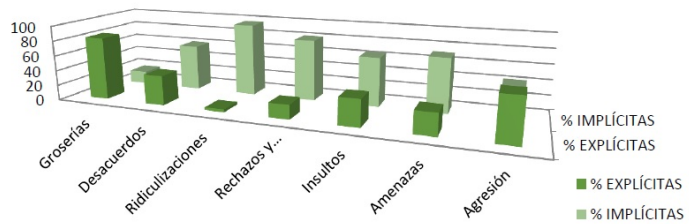
La teoría de la cortesía muestra cómo los actos de habla pueden –de entre otras acciones– eliminar lazos sociales; por tanto, confirma la idea de violencia en el lenguaje que se plasma en la teoría de la descortesía.

Teniendo en cuenta que la bibliografía sobre violencia lingüística no presenta un criterio uniforme para su clasificación, hemos propuesto nuestra propia clasificación de actos violentos. Esta clasificación incluye diferentes tipos de actos violentos según el grado de violencia, considerando el ataque a la imagen del emisor y del interlocutor.

La teoría de la relevancia se ha empleado para sustentar la afirmación de que las categorías de nuestra clasificación pueden realizarse de forma explícita e implícita.

Se aprecia que, salvo en la categoría de *groserías*, las demás categorías propuestas recurren al significado implícito para transmitir el contenido violento.

- La categoría que presenta mayor contraste entre explícito e implícito es la de *ridiculizaciones*. Son generalmente comparativas y, en la mayoría de casos, humorísticas (96%).



	Groserías	Desacuerdos	Ridiculizaciones	Rechazos y Discriminaciones	Insultos	Amenazas	Agresión
% EXPLÍCITAS	83	39	4	19	35	29	57
% IMPLÍCITAS	17	61	96	81	65	71	43

Fig. 3. Categorías de Violencia. Explicitud e Implicitud.

- La segunda categoría con mayor porcentaje de ejemplos implícitos es la de *rechazos y discriminaciones* (81%); éstos apelan al conocimiento histórico, como en el caso de la monarquía y el racismo, y al conocimiento del mundo en general, como en el caso del machismo.
- En los *insultos*, pese a la mayor presencia de formas explícitas respecto a las *ridiculizaciones*, abundan los enunciados con violencia implícita que se sirven de las implicaturas contextuales; aunque aparezcan insultos marcados como *imbécil* o *gilipollas* (35%), son más comunes los enunciados como *Se te ha caído el cerebro* (65%). Lo implícito de los *insultos* se puede deber a que los casos extraídos pertenecen a grupos (en el caso de Facebook) o bien a hashtags con gran número de tuits; por tanto, la generalización o el uso de implicaturas contextuales distancia al emisor del enunciado, lo que atenúa el insulto directo al interlocutor (en estos casos es desconocido). Las generalizaciones también tienen fuerte presencia en los *insultos* (15%), lo que incrementa el número de casos implícitos.
- Los *desacuerdos o desaprobaciones* no muestran tanto contraste como las otras categorías (39% explícitos y 61% de implícitos); sin embargo, aparecen frecuentemente exclamaciones e interrogaciones retóricas.
- En las *amenazas* también se observa una diferencia sustancial entre casos directos e indirectos (31% explícitas y 69% implícitas): los enunciados con violencia implícita expresan el deseo de realizar la acción violenta en un futuro, pero no se explicita, como en la fórmula marcada de *amenaza* (X va a *acción violenta* a Y).

- Las *agresiones* es la categoría con menor contraste: posee un total de 43% de casos implícitos. Aunque en gran parte de los casos se trata de enunciados que apoyan acciones violentas, la agresión aparece explicitada en más de la mitad de los enunciados analizados (57%).
- Las *groserías*, por último, a causa de su forma generalmente marcada, presenta el caso inverso: mayor número de ejemplos explícitos (83%).

Brown y Levinson (1987) y Culpeper (1996) presentan en sus teorías diferentes estrategias lingüísticas de cortesía, por un lado, y de descortesía, por otro lado. En la propuesta realizada en este trabajo se han examinado las estrategias de las distintas categorías de violencia lingüística, mediante las cuales se realiza el acto verbal. Tras el análisis del corpus, se observa que los recursos más empleados en las diferentes categorías son dispares, como muestra la Tabla 1.

Cada categoría posee tres recursos, pues se ha tomado la distinción realizada por Culpeper (1996) de gradación de explícito a implícito: inferencia por la forma, inferencia por el contexto e inferencia por convención.

Como muestra la Tabla 1:

- Conforme aumenta el grado de violencia lingüística, aumenta el grado de implícito de los enunciados: las categorías de *groserías* o *desacuerdos y desaprobaciones* se sirven de implicaturas contextuales, y las *amenazas* apelan al conocimiento compartido. El caso de la *agresión* es particular (es necesario dar indicios para referirse a una agresión física), pero también, en varios enunciados, se alude a noticias de agresiones físicas (por ejemplo, el puñetazo a Rajoy).
- En los *insultos*, debido a su componente cultural, es lógico que la estrategia de insulto implícito sea la generalización o insultos inferidos por el contexto lingüístico.
- Es particular el caso de las *ridiculizaciones*, la categoría que presenta mayor caso de números implícitos, cuyo recurso más numeroso es la comparación, con personajes o acciones de actualidad, y, concretamente, el empleo de enunciados de eco.

Teniendo en cuenta que determinados actos de habla amenazan a la imagen y que los actos de habla descorteses tienen la intención de agredir dicha imagen, parece lógico que la violencia lingüística esté presente en estas plataformas. La figura 4 muestra la presencia de las diversas categorías de violencia lingüística en las redes sociales, en general, y en el corpus analizado, en concreto.

Categoría	Explícito			Implícito
	Inferido por la forma	Inferido por el contexto	Inferido por convención	
Groserías	Palabras malsonantes	Referencia al asco y los efectos que provoca	Alusión a acciones escatológicas o sexuales	
Desacuerdos y Desaprobaciones	Expresión de desaprobación	Sensación de asco y vergüenza	Interrogaciones y exclamaciones	
Ridiculizaciones	Risa	Comparaciones	Enunciados de eco, enunciados infantiles	
Rechazos	Exhortación	Oraciones sarcásticas	Expresión de hartazgo, enunciados de eco	
Insultos	Disfemismos	Generalizaciones	Acusación de inconsciencia	
Amenazas	Amenazar	Explicitación del deseo de dañar físicamente al interlocutor	Referencia a actos violentos amenazantes	
Agresión	Alusión a agresión física	Implicación no directa del emisor en el acto violento	Exaltación de agresiones físicas	

Tabla 1. Categorías de Violencia y Estrategias.

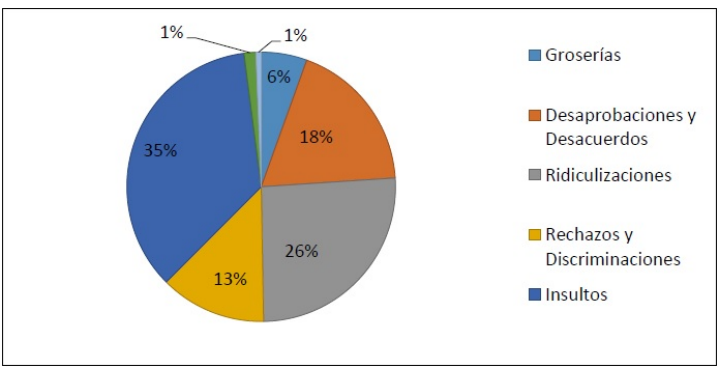


Fig. 4. Categorías de Violencia y su Presencia.

Como se ha explicado, el *insulto* y la *amenaza* suponen el mismo grado de ataque, según si se ataca a la imagen positiva o a la negativa. Por ello, se han tomado por separado en el análisis del corpus.

En vista de los resultados que muestra la figura 4, se puede afirmar que la violencia lingüística está presente en diferentes grados; sin embargo, dicha presencia es desigual. Los resultados del análisis muestran que:

- Las categorías de *insultos* y *ridiculizaciones* son las más presentes en el corpus analizado (35% y 26%, respectivamente). La afirmación de que, en las redes sociales, los fines que se persiguen con la violencia lingüística son la agresión, por un lado, y el alarde de ingenio, por otro (Díaz Pérez 2014), se confirma.
- Las *desaprobaciones o desacuerdos* ocupan el tercer lugar (18%), hecho lógico a causa del uso de estas plataformas virtuales como medio de expresión –teniendo en cuenta las temáticas seleccionadas como fuente del corpus–. Sin embargo, esta libertad de expresión generalizada a veces se confunde con la falta de respeto y educación, que queda demostrada en los *rechazos y discriminaciones* (13%). Las características del medio –anonimato, distancia temporal y espacial, escasa información contextual–, por un lado, y las características de las redes sociales –plataforma social, imagen, perfil "manipulado", interconexión entre varios usuarios simultáneamente–, por otro lado, favorecen este paso del desacuerdo al rechazo.
- Por último, las *amenazas* y las *groserías*, debido, quizá, a su esencial explicitud, tienen presencia escasa en el corpus analizado (1% y 6%, respectivamente). Cabe mencionar que las groserías, en muchos casos (en un 31% del corpus), son empleadas como recurso expresivo, por lo que refuerzan la violencia. El caso de la *agresión* es escaso (1%), aunque presente, dato de especial interés teniendo en cuenta el no conocimiento entre los usuarios emisores del corpus de ejemplos.

5.2 Conclusiones

Tras el análisis y el estudio realizado, las conclusiones a las que llegamos son las siguientes:

La violencia lingüística no es una, sino que comprende varias categorías graduables, pero todas ellas son violentas y, por tanto, atacan al interlocutor. La percepción de que únicamente los grados máximos de violencia (insultos, amenazas y agresión) son violencia propiamente dicha es, como queda demostrado, errónea; son varias las categorías que dañan la imagen del interlocutor, aunque el ataque sea de diferente grado. De hecho, el Código Penal condena el insulto y la amenaza (véase *Código Penal*, artículos 169, 170 y 171 para las amenazas, y artículo 208, para la injuria),

la discriminación, y la incitación al odio y la violencia contra grupos o asociaciones por motivos racistas, ideológicos o de enfermedad (artículo 510).

Los recursos que se emplean para hablar violentamente son diversos y, en la mayoría de casos, apelan al conocimiento histórico, cultural y tecnológico. Cada categoría se materializa tanto explícita (inferido por la forma) como implícitamente (inferido por el contexto o por convención). En la mayoría de casos, la violencia lingüística pasa por materializarse de forma implícita. Sin embargo, si la prueba no es explícita, es decir, si el acto no es directo, no se considera delito. Teniendo en cuenta que es mucha más la violencia implícita que la explícita, cabría replantearse el considerar solamente los casos explícitos como actos de violencia, así como la necesidad de presentar pruebas explícitas para poder denunciar (en el caso del maltrato psicológico, por ejemplo).

En relación a las redes sociales, estas suponen un nuevo modo de comunicación, como también de relaciones sociales (poseen diferencias con las presenciales); por ende, se observa un nuevo contexto en el que aparece la violencia lingüística. Se observa una gran presencia de ataque a través de *insultos* y *amenazas*, de humor por medio de *ridiculizaciones* y de crítica mediante *desacuerdo y desaprobaciones*. Teniendo en cuenta que la comunicación virtual es en gran parte lingüística y, particularmente, "escrita", queda demostrado el importante peso de la violencia lingüística, poco considerada socialmente respecto a la violencia física.

Al tratarse de plataformas sociales, en las que cada usuario posee una imagen "virtual", se propicia el alarde de ingenio y, por tanto, los enunciados implícitos. Las *ridiculizaciones*, categoría "intermedia" en la clasificación propuesta, suponen una burla en presencia de miles de usuarios, lo que propicia el *ciberacoso*, pues, además, el testimonio escrito de la ridiculización "alarga" el ataque. También es necesario un planteamiento estricto que limite tanto las libertades de expresión como el respeto a los usuarios, sobre todo teniendo en cuenta el auge de las redes sociales entre los adolescentes.

De especial interés es el hecho de que, pese al anonimato y la distancia social que presenta el mundo virtual, la violencia lingüística es en la mayor parte de los casos implícita; si en las relaciones virtuales, con mayor distancia y desconocimiento entre interlocutores, tiene más presencia la violencia lingüística en su vertiente implícita, es lógico que en las relaciones sociales cara a cara, dicha presencia sea todavía mayor. Todo ello lleva a reafirmar la necesidad de reflexionar sobre la explicitud requerida para considerar los casos de violencia como tales.

En síntesis, el estudio y el análisis realizado demuestran que la violencia lingüística tiene el mismo peso que la violencia física; se sirve de otra forma de materialización,

pero la acción es la misma: *violencia*. En muchos casos es implícita y puede no ser percibida claramente, lo que favorece su incremento, y las redes sociales suponen un nuevo contexto en el que la violencia está presente y, en concreto, la lingüística. Queda demostrada la necesidad de cuestionar y observar los "vacíos legales" de estas nuevas formas de relaciones sociales.

References

1. Albelda Marco, M. & Barros García, M.J. (2013). *La cortesía en la comunicación*. Madrid: Arco Libros.
2. Androutsoupoulos, J. & Beisswenger, M. (2008). Introduction. Data and Methods in Computer-Mediated Discourse Analysis. *Language@Internet*, 5(9): 1–7.
3. Bañón Hernández, A.M. (2010). Comunicación destructiva y agresividad verbal en la Red. Notas en torno al discurso sobre las enfermedades raras. *Discurso & Sociedad*, 4(4): 649–673.
4. Bell, A. (1984). Language style as audience design. *Language in Society*, 13: 145–204.
5. Bernal Linnersand, M. (2007). *Categorización sociopragmática de la cortesía y de la descortesía: Un estudio de la conversación coloquial española*. Estocolmo: Stockholm University.
6. Blas Arroyo, J.L. (2008). La descortesía verbal en contextos institucionales: entre la realidad y el espectáculo. In Fuentes Rodríguez, C. & Alcaide Lara, E.R. (eds.), *Manifestaciones textuales de la descortesía y agresividad verbal en diversos ámbitos comunicativos* (pp. 78–97). Sevilla: Universidad Internacional de Andalucía.
7. Blas Arroyo, J.L. (2010). La descortesía en contextos de telerrealidad mediática. Análisis de un corpus español. In Orletti, F. & Mariottini, L. (eds.), *(Des)cortesía en español. Espacios teóricos y metodológicos para su estudio* (pp. 183–208). Roma-Estocolmo: Università degli Studi Roma Tre-EDICE.
8. Bornhauser, N. (2012). Consideraciones en torno a la violencia. "Más allá" de las palabras. *Atenea*, 5062: 137–152.
9. Bousfield, D. & Locher, M.A. (2008). *Impoliteness in language: studies on its interplay with power in theory and practice*. Alemania: Deutsche Nationalbibiliothek.
10. Bravo, D. (1999). ¿Imagen "positiva" vs. Imagen "negativa"? Pragmática sociocultural y componentes de face. *Oralia*, 2: 155–184.
11. Bravo, D. (2008). Situación de habla, recursos comunicativos y factores lingüísticos en la interpretación de objetivos de cortesía. In Briz, A., Hidalgo, A., Albelda, M., Contreras, J. & Hernández, N. (eds.), *Cortesía y conversación: de lo escrito a lo oral. III Coloquio internacional programa EDICE* (pp. 12–24). Valencia: Universitat de València.
12. Brenes Peña, M.E. (2007). Estrategias descorteses y agresivas en la figura del tertuliano televisivo: ¿trasgresión o norma? *Lingüística en la red*: 1–19

13. Brenes Peña, M.E. (2009a). La agresividad como espectáculo en la televisión de hoy. La entrevista televisiva. In Fuentes Rodríguez, C. & Alcaide Lara, E.R. (eds.), *Manifestaciones textuales de la descortesía y agresividad verbal en diversos ámbitos comunicativos* (pp. 141–160). Sevilla: Universidad Internacional de Andalucía.
14. Brenes Peña, M.E. (2009b). *La agresividad verbal y sus mecanismos de expresión en el español actual*. Universidad de Sevilla, Sevilla.
15. Brenes Peña, M.E. (2009c). La expresión de la amenaza en el lenguaje juvenil. *Cultura, lenguaje y representación*, 7: 39–58.
16. Campos Santelices, A. (2010). *Violencia social*. Costa Rica: Editorial Universidad Estatal a Distancia.
17. Carrasco Santana, A. (1999). Revisión y evaluación del modelo de cortesía de Brown & Levinson. *Pragmalingüística*, 7: 1–44.
18. Cepeda, G. (2007). Discurso y (des)cortesía. *Discurso & Sociedad*, 1(2): 248–269.
19. Ciber Sociedad (2009). IV Congreso de la Ciber Sociedad 2009, Crisis analógica, futuro digital. *Redes sociales y medios de comunicación*.
20. Colín Rodea, M. (2011). Selecciones lingüísticas y discursivas en la "visibilización" de la dimensión moral de la violencia: el insulto moral. *Discurso & Sociedad*, 5 (3): 422–468.
21. Crystal, D. (2002). *El lenguaje e Internet*. Madrid: Cambridge UniversityPress.
22. Culpeper, J. (1996). Toward an Anatomy of Impoliteness. *Journal of Pragmatics*, 25: 349–67.
23. Culpeper, J. (2011). *Impoliteness: using language to cause offence*. Cambridge: Lancaster University.
24. Culpeper, J. (2016). Impoliteness strategies. In Capone, A. & Mey, J.L (eds.), Interdisciplinary Studies in Pragmatics, Culture and Society.Perspectives. *Pragmatics, Philosophy & Psychology*, 4: 421–445.
25. Díaz Pérez, J.C. (2014). Creatividad léxica y descortesía en los medios de comunicación virtual. *Revista de Filología*, 32: 81–98.
26. Escandell Vidal, V. (1993). *Introducción a la Pragmática*. Barcelona: Ariel.
27. Fant, L. (1989). Cultural mismatch in conversation: Spanish and Scandinavian communicative behaviour in negotiation settings. *Hermes Journal of Linguistics*, 3: 251–260.
28. Fuentes, C. (2010). Violencia y discurso. *Discurso & Sociedad*, 4(4): 643–648.
29. Fuentes Rodríguez, C. & Alcaide Lara, E.R. (2008). *(Des)cortesía, agresividad y violencia verbal en la sociedad actual*. Sevilla: Universidad Internacional de Andalucía.
30. Fuentes Rodríguez, C. & Alcaide Lara, E.R. (2009). *Manifestaciones textuales de la descortesía y agresividad verbal en diversos ámbitos comunicativos*. Sevilla: Universidad Internacional de Andalucía.
31. Garrido Rodríguez, C. (2006). Procedimientos para la cortesía en la conversación. *Cultura, lenguaje y representación*, 3: 57–71.
32. Goffman, E. (1967). *Interaction Ritual: Essays on Face-to-Face Behaviour*. Nueva York: Doubleday Anchor Books.
33. Gómez Arraigada, H. F. (2013). Desinformación en Internet y hegemonía en redes sociales. *Revista GPT Gestión de las Personas y Tecnología*, 16: 28–34.

34. Grice, H.P. (1975). Logic and Conversation. In Davis, S. (ed.) (1991), *Pragmatics. A Reader* (pp. 305–315). New York: Oxford University Press.
35. Haverkate, H. (1994). *La Cortesía Verbal*. Madrid: Gredos.
36. Haverkate, H. (2012). Cortesía y veracidad. *Lingüística y Literatura*, 62: 19–27.
37. Iglesias Moreno, P.A. (2012). El lenguaje de las Redes Sociales. In De Haro, M.V., Gandío, M.M. & Hernández, M. (coords.), *Historias en red. Impacto de las redes sociales en los procesos de comunicación*. Murcia: Ediciones de la Universidad de Murcia.
38. Kaul de Marlangeon, S. (2008). Tipología del comportamiento verbal descortés en español. In Briz, A.M., Hidalgo, A., Albelda, M.; Contreras, J. & Hernández, N., *Cortesía y conversación: de lo escrito a lo oral. III Coloquio internacional programa EDICE* (pp. 254–266). Valencia: Universitat de València.
39. Kaul de Marlangeon, S. (2012). Encuadre de aspectos teórico-metodológicos de la descortesía verbal en español. In Escamilla Morales, J. & Henry Vega, G. (eds.), *Miradas multidisciplinares a los fenómenos de cortesía y descortesía en el mundo hispánico* (pp. 76–107). Barranquilla-Estocolmo: Programa EDICE.
40. Kolakowski, L. (1999). De la violencia. *Discurso & Sociedad*, 5(3): 422–468.
41. Lambert Graham, S. (2008). A manual for (im)politeness?: The impact of the FAQ in an electronic community of practice. In Bousfield, D. & Locher, M.A (eds.), *Impoliteness in language: studies on its interplay with power in theory and practice* (pp. 281–304). Alemania: Deutsche Nationalbibiliothek.
42. Mancera Rueda, A. (2015). Los estudios sobre (des)cortesía y actividades de imagen en las redes sociales: notas para un estado de la cuestión. *Textos en Proceso*, 1(1): 50–70.
43. Medina López, J. (2011). Modelos de (des)cortesía verbal en la prensa española, el caso de "El Día" (Tenerife). *RILCE. Revista de Filología Hispánica*, 29(2): 365–388.
44. Ministerio de Justicia (2016). *Código penal y legislación complementaria*. Edición actualizada a 21 de enero de 2016, Boletín Oficial del Estado, Agencia Estatal Boletín Oficial del Estado, Madrid.
45. Pano Alamán, A. & Mancera Rueda, A. (2014a). Identidades falsas en Twitter: la ironía y el humor verbal como mecanismos paródicos. *Discurso y sociedad*, 8(3): 507–536.
46. Pano Alamán, A. & Mancera Rueda, A. (2014b). Las redes sociales como corpus de estudio para el Análisis del discurso mediado por ordenador. *Humanidades digitales: desafíos, logros y perspectivas de futuro*, Anexo 1: 305–315
47. Pano Alamán, A. & Moya Muñoz, P. (2016). Una aproximación a los estudios sobre el discurso mediado por ordenador en lengua española. *Tonos Digital, Revista Electrónica de Estudios Filológicos*, 30.
48. Portolés, J. (2004). *Pragmática para hispanistas*. Madrid: Síntesis.
49. Rius, M. ¿Somos malhablados?, *La Vanguardia, Vida*. Recuperado el 20/01/2012 de http://www.lavanguardia.com/estilos-de-vida/20120120/54244529265/somos-malhablados.html
50. Ruiz Antón, L.F. (1998). *La acción como elemento del delito y la teoría de los actos de habla: cometer delitos con palabras*. Madrid: Universidad Complutense de Madrid.
51. Searle, J. (1976). Actos de habla indirectos. Una taxonomía de los Actos Ilocucionarios. *Teorema*, 6(1): 43–77.

52. Searle, J. (1986). *Actos de habla*. Madrid: Cátedra.
53. Sperber, D. & Wilson, D. (1986). *La Relevancia. Comunicación y procesos cognitivos*. Madrid: Visor.
54. Terkourafi, M. (2008).Toward a unified theory of politeness, impoliteness, and rudeness. In Bousfield, D. & Locher, M.A. (eds.), *Impoliteness in language: studies on its interplay with power in theory and practice* (pp. 45–76). Alemania: Deutsche Nationalbibiliothek.
55. Vivas Márquez, J. & Ridao Rodrigo, S. (2015). "Lo siento pero me parecen horribles!!!" Análisis pragmalingüístico de la descortesía en la red social Facebook. *Revista de Filología*, 33: 217–236.
56. Yus, F. (1997). *Cooperación y relevancia. Dos aproximaciones pragmáticas a la interpretación*. Alicante: Publicaciones de la Universidad de Alicante.
57. Yus, F. (2010). *Ciberpragmática 2.0.* Barcelona: Ariel.

Author's Biodata

Susana Campillo Muñoz es graduada en Lengua y Literatura Hispánicas por la Universitat Rovira i Virgili (Tarragona).

TRIANGLE

Language, Literature and Computation
Lenguaje, Literatura y Computación
Llenguatge, Literatura i Computació

FOCUS AND SCOPE: Triangle is an open access electronic publication devoted to any topic related to language, literature and computation. Triangle is addressed to a specialized audience interested in the adoption of new methods (especially formal and computational) in the study of language and literature.
Triangle was born in 2010 with the aim of promoting interdisciplinary research in the fields of linguistics, literary studies and computer science. Triangle intends to disseminate original, relevant, innovative and high-quality papers in the space of confluence of those three research areas.
The journal name, Triangle, intends to reflect, through the three vertices of the figure, the three topics of interest for this publication (language, literature and computation) and aims to evidence that preference is given to articles that focus on the edges of the triangle, this is, works that relate, at least, two of the three involved disciplines.
Triangle welcomes contributions in any of the three journal areas: linguistics, literary studies and computer science. However, preference is given to articles involving/relating at least two of the vertices of the triangle. Studies relating language and literature or any of these two disciplines with computation will be warmly received.
Contributions to Triangle may be made in any of the following categories/forms:

- **Papers** that report significant new research in any of the journal topics.
- **Monographic Issues** containing the contribution of a single author that focuses on new research results or that presents didactic introductions to new research topics or that surveys the state of the art of an area in order to provide a good starting point for non-experts.
- **Special Issues** containing a collection of essays/articles, written by several researchers, around a well-established topic.

LANGUAGES: The languages of submission and publication are preferably English and Spanish. However, papers in Catalan and French will be also accepted.

PEER REVIEW PROCESS: Triangle is a peer-reviewed journal and follows a double-blind refereeing procedure.

EVALUATION CRITERIA: No particular theories or scientific trends are favoured. Scientific quality and scholarly standing are the only criteria applied in the selection of papers accepted for publication.

Universitat Rovira i Virgili

Este volumen incluye dos trabajos. El primero propone una metodología que permite inducir la gramática del yorùbá a partir de la gramática del inglés utilizando un modelo basado en restricciones, las *Womb Grammars*. El segundo se centra en el estudio de la violencia lingüística y, a partir de un corpus de ejemplos extraídos de Twitter y Facebook, analiza los grados de violencia y los recursos y las técnicas que se utilizan para emitir actos de habla violentos.

This volume includes two works. The first one presents a methodology for inducing the grammar of Yorùbá from the grammar of English by using a constraint solving model, the Womb Grammars. The second paper focuses on the study of linguistic violence and, from a corpus of interactions extracted from Twitter and Facebook, analyzes the degrees of violence in speech acts and the resources and techniques used to produce violent speech acts.

UNIVERSITAT ROVIRA i VIRGILI
Departament de Filologies Romàniques